nature

THE INTERNATI WEEKLY JOURNAL OF SCIENCE

UK BIOBANK

Genetic and health data from half a million people in the United Kingdom

PAGES 194, 203 & 210

TECHNOLOGY

TIME TO THINK SMALL

Fleets of tiny satellites could change space exploration

PAGE 185

OPTOELECTRONICS

TURNING UP THE LIGHT

Boost in performance for perovskite LEDs

PAGES 197, 245 & 249

DEVELOPMENTAL BIOLOGY

TWO WAYS TO GROW

A second source for the cells that line blood vessels

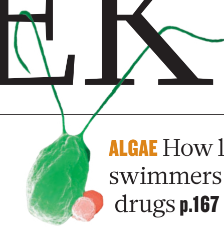
PAGES 195 & 223

NATURE.COM

11 October 2018

Vol. 562, No. 7726

THIS WEEK



EDITORIALS

NOBELS Steps to make premier prizes more representative p.164

ICE FLASH The cold keys to material superconductivity p.166

ALGAE How little green swimmers could deliver drugs p.167

A matter of (half) degrees

The latest IPCC assessment on a 1.5 °C increase makes it clear that there is no safe level of global warming. But will people listen?

Readers who remember the 1960s and 1970s have already witnessed something remarkable in Earth's shared history: roughly half a degree's worth of global warming. And, yes, science now confirms the often-expressed sentiment that something feels different. More-intense heatwaves; more-powerful storms; more wildfires. And more on the way.

The likely changes associated with another half degree of warming over the next few decades are discussed in the latest assessment by the Intergovernmental Panel on Climate Change (IPCC). The picture is gloomy. Policymakers and others must take from it a sense of urgency, an understanding that climate change is a problem for the here and now, and a conviction that they can make a difference.

The special report on 1.5 °C has its origins in the 2015 Paris climate agreement, in which 195 governments committed to limit global warming to "well below 2 °C" while "pursuing efforts to limit the temperature increase to 1.5 °C". Although their commitments to reduce emissions fall well short of either goal, governments still called on the IPCC to prepare a special report on the impacts that could be expected at 1.5 °C — and how much worse things would get if the temperature rise reached 2 °C (see page 172).

As the summary released on 8 October makes clear, 1.5 °C is troubling enough — but there is a world of difference between 1.5 and 2 °C. Yes, 1.5 °C would bring increases in troublesome weather, such as the heatwaves, droughts, storms and flooding. Deeper issues lurk: the planet is undergoing rapid changes in how it looks and functions, and as greenhouse-gas emissions rise, so, too, does the risk of permanent damage.

The Arctic Ocean is projected to be completely free of ice once per century with a 1.5 °C rise, or once per decade at 2 °C. Sea levels are set to continue rising well beyond 2100. Many of today's ecosystems will shift or disappear: literature covering 105,000 species suggests that 6% of insects, 8% of plants and 4% of vertebrates could lose half of their territory with even 1.5 degrees of warming; those numbers increase by two or three times in the case of 2 degrees. The situation may be even worse in the oceans. At 1.5 °C, the world could lose 70–90% of its coral reefs. They pretty much disappear entirely at 2 °C — a threshold beyond which the risk of irreversible loss of marine ecosystems increases dramatically.

Governments also asked the IPCC for more information about what it would take to halt global warming at 1.5 °C. Although earlier estimates suggested that the world could blow through its 1.5 °C carbon budget within several years, the new budgets allow for a steady — but dramatic — downward trajectory that ends with zero carbon emissions in the middle of this century. Recent research does suggest the world has a bit more breathing space for reducing emissions to meet that goal.

But there is a danger that this signal — that we have more time than we thought — becomes the take-home message for policymakers. That would be a mistake. First, the carbon budgets are based on relatively recent and still-controversial research, and could yet be revised. Second, as the IPCC report makes clear, going carbon-neutral by mid-century is a terribly daunting challenge. Modelled scenarios that maintain warming at 1.5 °C assume that renewable energy sources such as wind and solar must account for 70–85% of global electricity production by 2050. Natural-gas-fired power plants equipped with carbon-capture and carbon-sequestration technology account for just 8% of the projected power needs, with coal close to zero.

This has dire implications for fossil-fuel infrastructure and investments, and will affect the price of energy, consumer products and jobs in many places. Governments — and businesses — will need to ensure that people who work in the fossil-fuel industries are not forgotten in the process. But the report also makes it clear that the benefits of aggressive action far outweigh the costs. Now in its 30th year, the IPCC has issued a valuable assessment based on a flurry of research conducted since 2015. It is just the latest in a long series of reports that now serve as both a scientific foundation and a warning about the perils of unchecked global warming. Unfortunately, the governments of the world have yet to take heed of this report's calls to spur new political momentum.

> ***"At 1.5 °C, the world could lose 70–90% of its coral reefs. They pretty much disappear entirely at 2 °C."***

Projections based on current emissions commitments suggest that the world is on track for around 3 °C of warming by the end of the century. On the basis of the cascade of changes now projected for 1.5 °C, that is a frightening prospect indeed. If those days of the 1960s and 1970s seem as if they are from a different world, it's because they are. ■

Crowd screen

Precision medicine relies on studies that track huge numbers of people.

Precision medicine aims to improve treatments for individuals, but to do so it needs information from crowds. Only by tracking the health of large numbers of people can the influence of genetics be teased out and incorporated into future tailored treatments. Scientists now report the success of such a project, the UK Biobank, which holds genetic, physical and clinical data from a large cohort of individuals in the United Kingdom. Many nations have launched biobank projects, including Estonia, Japan, Canada and Finland. Iceland was a pioneer, but the United Kingdom has gone much larger: by 2010, the UK Biobank had a prospective cohort of some 500,000 individuals, aged 40–69 at recruitment. Following this age group enables a focus on diseases of middle age and later.

In this week's *Nature*, researchers report the first descriptions of the full cohort, including genome-wide genetic data for all individuals

(see C. Bycroft *et al. Nature* **562**, 203–209; 2018). In a second study, researchers report brain imaging of 10,000 individuals, which reveals genetic influences on brain structure and function, and shows correlations with neurodegenerative, psychiatric and personality traits (see L. T. Elliott *et al. Nature* **562**, 210–216; 2018). Such findings are invaluable, but the usefulness of the UK Biobank project goes beyond immediate clinical relevance. It offers lessons for researchers establishing population-cohort and genomic-medicine projects elsewhere.

The UK Biobank benefits greatly from the infrastructure and centralization of the United Kingdom's National Health Service (NHS). In addition to recruitment through NHS centres, the project follows participants by accessing their health records and national registries, including those for deaths and cancer diagnoses.

Notably, the UK Biobank is the first project to demonstrate the successful collection and sharing of linked genetic, physical and clinical information on a population scale. All involved should thank the 500,000 volunteers across the United Kingdom who responded to their invitations and agreed to contribute their time, samples and health information. Buoyed by this success, UK Health Secretary Matt Hancock last week confirmed a significant expansion of genomic medicine in the NHS, which will grow the 100,000 Genomes Project to sequence the genomes of 1 million people through the NHS and the UK Biobank. This is part of an even more ambitious project to sequence up to 5 million genomes over the next 5 years, including those of seriously ill children and people with rare types of cancer.

Such scale is important, but so is diversity. The UK Biobank is filled with people who lived near an assessment centre and agreed to participate. Aiming for a more diverse population is an additional challenge, but a worthy one. The All of Us cohort study in the United States is making efforts to do this with targeted recruitment.

In many population-cohort studies, the data are not made accessible to other researchers until the initial findings are published, and even then only a few make their full data sets available. The UK Biobank, funded primarily by the Medical Research Council and the Wellcome Trust, and run as a charity, has taken an important stand. It has generously made its full data sets, as well as all results from studies conducted by researchers using these data, available from the outset.

"That is the future of medicine: wisdom from crowds."

The value of such an open approach is clear. Since the UK Biobank opened general access to its database in March 2012, there have been at least 8,294 approved registrations, and 796 formally registered projects are under way. The results of these studies have been communicated in more than 500 publications in peer-reviewed journals and in over 100 preprints on a dedicated bioRxiv channel.

In particular, this access has allowed researchers to quickly search for genetic associations for a large and diverse collection of clinically relevant traits. A News Feature on page 181 explores what we have learnt from these larger-scale studies about genetic risk of disease, particularly the development of risk scores involving multiple genes, which could help to guide preventive measures for some common ailments such as coronary artery disease. Although controversial, such tests are already being developed commercially.

Many of these studies have aggregated UK Biobank data with other data sets to enable studies on a much larger scale, some reaching more than 1 million individuals. That is the future of medicine: wisdom from crowds. ■

Noble effort

The bodies that govern the Nobel prizes must do more to achieve equality.

Newly minted Nobel laureates Donna Strickland and Frances Arnold are outstanding scientists. They are also women. Advocates of equality in science understandably feel torn between celebrating these women's achievements and shouting that their triumph does not mean the problem of equality in science is solved. The day when attributes such as the gender, sexuality or ethnicity of a Nobel prizewinner is not relevant will be a great one, but it is not today. The bodies that govern the Nobel prizes must to do more to achieve that.

Gender is an area in which the Nobel skew is particularly obvious. An abysmally small number of women have been awarded one of the science Nobel prizes — this year's awards bring the tally to 19 women out of 607 laureates (including people who have won two science Nobels) — still just 3%, or 9% over the past decade. This is important because, for good or ill, the Nobel prizes matter. Laureates become science superstars and role models whose voices are amplified overnight. The prizes signal to the public who is the best of the best. Many of them recognize work from a time when the representation of women and people of colour was even lower than it is today. But, crucially, the awards are part of a system in which the balance remains tipped in the favour of Western white men, not just a product of that system.

Commendably, the bodies in Stockholm that award the prizes — the Royal Swedish Academy of Sciences for chemistry and physics, and the Nobel Assembly at the Karolinska Institute for physiology or medicine — have recognized that there are more people from under-represented groups who deserve the prize than receive it. And changes that the bodies introduced this year, which will take effect in 2019, could help. (These include flagging to nominators that they can select multiple candidates, and that they should consider diversity in gender, geography and topic — both of which encourage nominators to look beyond their immediate biases.)

But if the committees receive too few women candidates, why not highlight this by publishing aggregate demographic data on nominations? Right now, we know nothing about whether two female science winners in 2018 is a blip or representative of shifting attitudes, nor at what stage of the process the problem really lies — in nominations or selections. Having data on any problem is the first stage in tackling it. Transparency is a growing movement in science — and rightly so. Similar efforts can be made for scientific prizes.

Instead, the Nobel committees hide such information behind statutes that say nominations must remain confidential for 50 years. A spokesperson for the Royal Swedish Academy of Sciences told *Nature* that this is to prevent attempts to interfere with the nomination process and to allow researchers to give their honest opinion on colleagues' work. But it is hard to see how revealing aggregate data threatens that. Moreover, Alfred Nobel's will, on which these rules are based, does not call for confidentiality — merely stipulating that "the prize be awarded to the worthiest person". In this case, changing the statutes or their interpretation to allow for greater transparency will only help to achieve that.

The committees should also look at their own diversity. And they should state robustly why deserving women rarely win: because biases that are baked into the scientific system subtly (and sometimes not so subtly) hinder their route to the top as well as their eventual recognition. Evidence squarely shows this (see page 165). The situation is compounded for scientists who are from sexual and gender minorities, who are people of colour, who are disabled or not from a Western country. The world is still waiting for the first black winner of a science Nobel.

The march of history is towards equality, and many more like Strickland and Arnold are no doubt waiting in the Nobel wings. These women's wins are sources of hope for those who come after them. Strickland says that she feels she has been treated with equality in her career, a point that should be celebrated. But when women winning is not unusual enough to provoke comment — that will be the day for true celebration. ■

SETH KATZ

Two Nobels for women — why so slow?

Women in science still don't get what they deserve, explains **Virginia Valian**, *20 years on from her landmark book on bias.*

Last week brought great news and irksome news for science. The total number of women to win science Nobel prizes grew from 17 to 19, with Frances Arnold's chemistry award for enzyme engineering and Donna Strickland's prize for laser physics. The last time a woman won the physics Nobel was in 1963; before that, it was 1903. At that rate, we would expect another around 2068.

My hope is that two female prizewinners in one year portends a faster pace for recognition of women's achievements in science.

Unfortunately, last week also brought a talk that shows how much further we have to go in appreciating women's contributions. At CERN, Europe's particle-physics laboratory near Geneva, Switzerland, Alessandro Strumia of the University of Pisa in Italy spoke at a session on women in physics. According to attendees (and my reading of his slides), Strumia asserted that women are given unfair advantages and yet are scarce owing to a lack of ability and lack of interest — claims that are controversial, at best. As evidence of discrimination against men, Strumia named a woman who was hired for a job he had also applied for; he suggested his qualifications had been stronger, because he had more citations. His talk also seemed to presume that citations are the only measure of scientific quality, and played down evidence that women are cited less often than men, even after controls for quality. CERN issued a statement describing the talk as "highly offensive", and said that it would suspend Strumia from CERN-affiliated activities pending an investigation. (Strumia maintains that his presentation was not sexist or biased.)

MEN AND WOMEN SLIGHTLY **OVERRATE** MEN'S PERFORMANCE AND ABILITIES AND SLIGHTLY **UNDERRATE WOMEN'S.**

I have spent the past 25 years studying the structural and psychological reasons for the paucity of women at the top of almost every field in academia, and I have written two books documenting data that show how women's careers are hindered. The first, *Why So Slow?*, appeared in 1998; the second, *An Inclusive Academy: Achieving Diversity and Excellence*, co-authored with psychologist Abigail Stewart of the University of Michigan in Ann Arbor, came out earlier this year. The second book shows that, despite improvements, progress is still slow.

To take just one example, an analysis of six disciplines at leading US universities in 2013 and 2014 found that men gave colloquia disproportionately more often than women did. The result held even after adjusting for the representation of women in each field (C. L. Nittrouer *et al. Proc. Natl Acad. Sci. USA* **115**, 104–108; 2018). It also found that women and men rated the importance of giving talks equally, and that they accepted invitations with similar frequency. Women don't choose not to talk. They simply aren't invited to do so as often as they should be.

Experiments and field studies find that both men and women slightly overrate men's performance and abilities and slightly underrate women's. The many instances in which women don't get their due — among others, being ignored in meetings, not being invited to peer-review research or being denied a Wikipedia entry (as happened with Strickland), a promotion or telescope time — add up. The accumulation of these disadvantages acts like compound interest, widening disparities over time.

In 2004, I gave a talk to an honours society at a City University of New York campus. I was the first female speaker since the series had begun in the 1960s. The next woman after me spoke in 2013. The lack of female speakers was not due to a dearth of options, or, I think, to any intention to discriminate. In fact, that is where most people go wrong. They mistake their intentions — to judge on merit — for fact. Scientists, who think that they are responsive to data, might be especially likely to mistakenly trust their own judgement as being unbiased.

In one study, researchers asked faculty members in chemistry, physics and biology departments to rate a CV for an applicant applying for a lab-manager position (C. A. Moss-Racusin *et al. Proc. Natl Acad. Sci. USA* **109**, 16474–16479; 2012). In general, faculty members were more likely to hire the lab manager if the CV was for a man (the team used 'John') than for a woman ('Jennifer'), despite the CV being identical in every other way. They were also more willing to mentor John than Jennifer, and to offer him a higher starting salary. The preference for the man was marked in those who thought that gender equity was not a problem.

When I reviewed Strumia's slides, I was perturbed by how much his talk ignored and oversimplified solid scientific work on sex and gender differences. I also saw that he gave short shrift to the large body of psychological, sociological and economic data that show how individuals and institutions put women (and under-represented groups, such as people of colour or those with disabilities) at a disadvantage. I would have expected more familiarity with scholarship that I and many others have documented.

Why am I discussing this backsliding talk in a happy week of two Nobel prizes for women? Because the talk matters: children, students, graduates, assistant professors and others develop an idea of what they can aspire to be in part by seeing who have become lecturers and principal investigators — as well as who wins prizes. We need to see a range of people. And we need evidence that people already there will accept us.

No field can afford to ignore or alienate half its potential contributors. If we want talent, we have to welcome it and nurture it, in all its diversity. In our book, Stewart and I describe policies that can make participation and recognition more fair, such as developing explicit criteria for identifying and evaluating candidates, rather than relying on flawed proxies, such as prestige. To do the best possible science, we need to bring out the best that people can offer. ■

Virginia Valian *is distinguished professor at Hunter College and the Graduate Center of the City University of New York.*
e-mail: vvalian@hunter.cuny.edu

SEVEN DAYS *The news in brief*

RESEARCH

Exomoon evidence

Astronomers have spotted what could be the first known moon to orbit an exoplanet, according to a study published on 3 October in *Science Advances*. The planet, dubbed Kepler-1625b, lies 2.4 kiloparsecs (about 8,000 light years) from Earth in the constellation Cygnus. Researchers have detected hints of a possible moon orbiting this planet before, in data from the Kepler space telescope. But now, on the basis of observations made with the much more powerful Hubble Space Telescope, researchers are a lot more confident that the exomoon is real. If confirmed, the discovery would mark a milestone in exploring planetary systems throughout the Galaxy. It would, among other things, allow scientists to test ideas of moon formation using examples from beyond the Solar System. See go.nature.com/2nvaiqo for more.

NASA

Hubble telescope stops collecting data

The Hubble Space Telescope stopped collecting science data on 5 October, because of a problem with one of the gyroscopes it uses to orient itself. Mission controllers expect to have Hubble working again soon. "Don't worry, Hubble has many great years of science ahead," says Kenneth Sembach, director of the Space Telescope Science Institute in Baltimore, Maryland, which operates the telescope. But the glitch underscores that Hubble, perhaps the most iconic space observatory in history, will eventually die. NASA astronauts cannot service the 28-year-old observatory as they once did, because the agency retired its space shuttles in 2011. On their last servicing mission in May 2009, astronauts replaced Hubble's six gyroscopes. Hubble can operate on just one gyroscope, but that limits its ability to point at targets.

POLICY

Arctic protections

Nine nations and the European Union signed an agreement on 3 October in Ilulissat, Greenland, to ban unregulated commercial fishing on the high seas of the central Arctic Ocean. Warming temperatures in recent years have resulted in open water during the summer in a large region of the usually frozen central Arctic. This has prompted concern from scientists and officials that commercial operations could enter an area of nearly 3 million square kilometres and deplete fish populations. The legally binding agreement prohibits fishing in the area until a regional fishery-management organization can be established to set scientifically based quotas and regulations. It also establishes scientific cooperation between the countries to share information about changing Arctic ecosystems and fish stocks. The nations involved are Canada, Denmark, Norway, Russia, the United States, China, Iceland, Japan and the Republic of Korea.

Gene-editing rules

Japan has issued draft guidelines that would allow the use of gene-editing tools in human embryos. The proposal was released on 28 September by an expert panel representing the country's health and science ministries. Although the country regulates the use of human embryos for research, there have until now been no specific guidelines on using tools such as CRISPR–Cas9 to make precise modifications to their DNA. Manipulating DNA in embryos could reveal insights into early human development. The draft guidelines will be open for public comment from next month and are likely to be implemented in the first half of next year. If adopted, the guidelines would restrict the manipulation of human embryos for reproduction, although this would not be legally binding. See go.nature.com/2gdrjlp for more.

BUSINESS

ResearchGate suit

Two journal publishers have launched legal proceedings in the United States against academic-networking site ResearchGate for copyright infringement. Elsevier and the American Chemical Society (ACS) say that the ResearchGate website violates US copyright law by making articles from their journals freely available. The two publishers filed the claim with the US District Court

for the District of Maryland on 2 October. ResearchGate, which is based in Berlin, declined to comment to *Nature*. In October 2017, the same publishers launched a similar suit for copyright infringement in Germany, and that case has not yet concluded. At the time, ResearchGate also declined to comment on this lawsuit. By the following month, ResearchGate had disabled public access to 1.7 million articles on its site. The Coalition for Responsible Sharing, a group of publishers — including Elsevier and the ACS — that formed to order ResearchGate to remove their papers from its site, estimates that up to four million copyrighted articles have been made available for free on the platform.

BBVA FOUNDATION FRONTIERS OF KNOWLEDGE AWARDS; YORICK JANSENS/BELGA VIA ZUMA

AWARDS

Economics Nobels

Two US economists, William Nordhaus (**pictured, left**) and Paul Romer (right), share the 2018 Nobel Prize in Economic Sciences for integrating climate change and technological change into macroeconomics. Nordhaus, at Yale University in New Haven, Connecticut, is the founding father of the study of climate-change economics. Economic models that he has developed since the 1990s are now widely used to weigh the costs and benefits of curbing greenhouse-gas emissions against those of inaction. His studies are central to determining the social cost of carbon — an attempt to quantify the total cost to society of greenhouse gases, including hidden factors such as extreme weather and lower crop yields. Romer, at the New York University Stern School of Business in New York, was honoured for his work on the role of technological change in economic growth. He is best known for his studies on how market forces and economic decisions facilitate technological change. His 'endogenous growth theory', developed in the 1990s, opened up avenues of research on how policies and regulations can prompt fresh ideas and economic innovation.

SPACE

Japan rover lands

A third rover touched down on the surface of asteroid Ryugu on 3 October, marking a hat-trick of successful landings for the Japanese Hayabusa2 space mission. The shoe-box-sized Mobile Asteroid Surface Scout (MASCOT) separated from the Hayabusa2 probe, which had moved temporarily to 51 metres from the asteroid's surface. The lander then descended to the asteroid in free fall. MASCOT is scheduled to visit three sites on the 880-metre-wide asteroid, using an external swinging arm to 'hop' around in the asteroid's low gravity. It is equipped to measure the temperature on the surface and during the descent, as well as the asteroid's magnetic field. It will also study the composition of the surface. Ryugu is made up of material from the early Solar System, and scientists think that studying the asteroid will give them an insight into the evolution of Earth and other planets. MASCOT was one of four landers aboard Hayabusa2. See go.nature.com/2imfcut for more.

ENVIRONMENT

Whaling rethink

Japan's fisheries agency says it will consider revising its sale of sei whale meat by February. The announcement came on 4 October, days after the Convention on International Trade in Endangered Species of Wild Fauna and Flora in Geneva, Switzerland, decided that selling products from sei whales (*Balaenoptera borealis*) violates restrictions on the sale of an endangered species. Although there has been a moratorium on whaling since 1986, Japan continues to hunt the mammals, mostly sei and minke, as part of what it calls a scientific research programme. It sells the meat and blubber, arguing that these products would otherwise be discarded. Last month, the International Whaling Commission, based in Cambridge, UK, rejected Japan's bid to restore commercial whaling.

TREND WATCH

An analysis of industry-funded clinical trials has found that drug companies are often heavily involved in the conduct and reporting of the research — but are not always transparent about it.

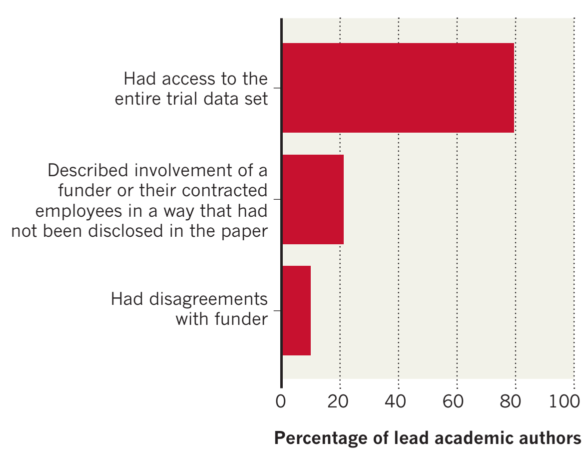
Kristine Rasmussen, a medical researcher at the Nordic Cochrane Centre in Copenhagen, and colleagues searched 7 high-impact medical journals, picking out the 200 most recent phase III and phase IV trials of drugs, vaccines and medical devices (K. Rasmussen *et al. Br. Med. J.* **363**, k3654; 2018).

They found that fewer than half of the trials had academics involved in data analysis, whereas 73% had funders involved. Rasmussen suggests that lack of time or statistical know-how could mean that many clinicians are happy to leave analysis to funders.

A survey of the trial's lead academic authors, completed by around 40%, found that only 79% of those that responded reported having access to the entire trial data, and 11% said that they had had disagreements with the funders. About 21% said a funder, or one of their contracted employees, had been involved in the research in a way that had not been declared in the paper.

INDUSTRY FUNDERS' INFLUENCE

How much say does industry have in the studies that it funds? A survey of 80 academic authors of industry-funded clinical trials revealed issues when publishing such work.



SOURCE: K. RASMUSSEN *ET AL. BR. MED. J.* **363**, K3654 (2018)

NEWS IN FOCUS

PICTURE ALLIANCE/AVALON.RED

People march in protest against Jair Bolsonaro, one of Brazil's presidential candidates, in the southern city of Curitiba.

POLITICS

Brazil's presidential election could savage its science

One leading candidate has proposed pulling the country out of the Paris climate agreement.

BY JEFF TOLLEFSON

A populist surge from a right-wing presidential candidate in Brazil that is threatening to upend the country's politics could have huge impacts on research budgets and environmental policies.

Jair Bolsonaro, a controversial politician often dubbed the 'Tropical Trump', has outlined plans that would weaken environmental protections and reorganize federal science programmes. He won the first round of voting on 7 October with 46% of the votes — just shy of the 50% he needed to avoid a run-off election.

Bolsonaro will face Fernando Haddad, a former São Paulo mayor who won 29% of the vote, in a run-off on 28 October. Haddad is the replacement candidate for former Brazilian president Luiz Inácio 'Lula' da Silva, a popular leader of the leftist Workers' Party who was barred from running in this election because he is in prison on corruption charges.

Years of economic woes and corruption scandals serve as a backdrop to the election. Brazil's federal science budget has declined sharply over nearly a decade, and pro-industry politicians are slowly chipping away at the country's environmental regulations. But the two leading presidential candidates have offered very different visions for addressing these issues, leaving scientists on edge.

Bolsonaro, a lawmaker from Rio de Janeiro in Brazil's lower house of Congress, often votes with the conservative rural caucus, which is actively seeking to weaken environmental regulations. He has proposed decentralizing federal science programmes — although it's unclear how he would do so — and merging environment ministry with the agriculture, livestock and supply ministry. Bosonaro has also suggested pulling Brazil out of the 2015 Paris climate accord.

In the Amazon region, scientists say, Bolsonaro is seeking to promote agricultural and industrial expansion at the expense of environmental protections and the rights of Indigenous communities.

The message to industry and agriculture seems to be that a Bolsonaro administration would let them do whatever they want in the Amazon, says Carlos Rittl, executive ▶

secretary of the Climate Observatory in São Paulo, a network of 37 groups focused on climate policy. If Bolsonaro won, it "would be a nightmare".

Bolsonaro — whose vice-presidential running mate has raised the spectre of military intervention to address political dysfunction — was once considered a long-shot candidate. The latest poll analysing run-off scenarios, however, shows Bolsonaro with a slight lead over Haddad.

"People say Bolsonaro stands no chance, but who knows," says Carlos Nobre, a climate scientist and former secretary for research and development policy at Brazil's Ministry of Science, Technology and Innovation.

BOOSTING SCIENCE

Haddad, by contrast, has a more mainstream vision for Brazil that emphasizes science, innovation and action on climate and environmental policies. He has promised to promote renewable energies, such as wind and solar, while fighting deforestation and maintaining protections for Indigenous territories in the Amazon.

And unlike Bolsonaro, who has called for more private-sector research and development, Haddad has committed to boosting federal spending on science. He has proposed raising the national investment in research and development to 2% of Brazil's gross domestic product, using government and private funding. That would bring the country's science spending in line with many industrialized nations.

It's unclear how feasible those spending goals are. One wrinkle is that in late 2016, Brazil adopted a constitutional amendment that caps government investments for 20 years, aside from adjustments for inflation.

Any policies that recognize and invest in science and technology are welcome, says theoretical physicist Luiz Davidovich, president of the Brazilian Academy of Sciences. He notes that, after adjusting for inflation, the science ministry's budget has decreased by roughly two-thirds since 2010, to around 3.4 billion reais (US$860 million).

Budget shortfalls have meant less money for equipment, federal grants, travel and postdoctoral fellowships for public-university researchers in Brazil. Despite this, Davidovich says, scientists are pressing on wherever possible.

Although science and technology factor in the campaigns of Bolsonaro and Haddad, it's too soon to tell what might happen after the election.

"The fact that they have science and technology in their programme does not mean it's going to be important when they become president," Davidovich says. "There is a big difference between what is written, and what is practised." ■

NASA/EYEVINE

Glaciers and sea ice won't be safe in a world that warms to 2 °C above pre-industrial levels.

GLOBAL WARMING

Clock ticking on climate action

IPCC sees small window to avoid worst effects of warming.

BY JEFF TOLLEFSON

Limiting global warming to 1.5 °C above pre-industrial levels would be a Herculean task, involving rapid, dramatic changes in how governments, industries and societies function, says the Intergovernmental Panel on Climate Change (IPCC). But even though the world has already warmed by 1 °C, humanity has 10–30 more years than scientists previously thought in which to kick its carbon habit.

To meet this target, the world would have to curb its carbon emissions by at least 49% of 2017 levels by 2030 and then achieve carbon neutrality by 2050, according to a summary of the latest IPCC report, released on 8 October. The report draws on research conducted since nations unveiled the 2015 Paris climate agreement, which seeks to curb greenhouse-gas emissions and limit global temperature increase to between 1.5 and 2 °C.

The world is on track for around 3 degrees of warming by the end of the century if it doesn't significantly reduce greenhouse-gas emissions. It could breach 1.5 °C between 2030 and 2052 if global warming continues at its current rate.

Scientists have "high confidence" that 1.5 °C of warming would result in a greater number of severe heat waves on land, especially in the tropics, the report says. They have "medium confidence" that there will be more extreme storms in areas such as high-elevation regions, eastern Asia and eastern North America. The risk of such severe weather would be even greater in a 2 °C world. Temperatures on extreme hot days in mid-latitudes could increase by 3 °C with 1.5 °C of global warming, or by 4 °C in a 2 °C world.

Two degrees of warming could destroy ecosystems on around 13% of the world's land area, increasing the risk of extinction for many insects, plants and animals. Holding warming to 1.5 °C would reduce that risk by half.

The Arctic could experience ice-free summers once every decade or two in a 2 °C world, versus once in a century at 1.5 °C. Coral reefs would almost entirely disappear with

2 degrees of warming, with just 10–30% of existing reefs surviving at 1.5 °C.

Without aggressive action, the world could become an almost impossible place for most people to live in, says Ove Hoegh-Guldberg, director of the Global Change Institute at the University of Queensland in St Lucia, Australia. "As we go toward the end of the century, we have to get this right."

IMPOSSIBLE DREAM

Given that current national commitments on greenhouse-gas emissions fall well short of the goals laid out in the Paris climate agreement, many scientists have argued that meeting even the 2 °C goal is almost impossible. But the IPCC report sidestepped questions of feasibility and focused instead on determining what governments, businesses and individuals would need to do to meet the 1.5 °C goal.

Measures include ramping up the installation of renewable-energy systems, such as wind and solar power, to provide 70–85% of the world's electricity by 2050, and expanding forests to increase their capacity to pull carbon dioxide from the atmosphere.

Most scenarios in the report suggest that the world would still need to extract massive amounts of carbon from the atmosphere and pump it underground in the latter half of this century. The technology to do this is in the early stages of development, and many researchers say that it could be difficult to develop it for use on a global scale.

Other proposed options involve changing lifestyles: eating less meat, riding bicycles more and flying less. The report also waded into murky questions about ethics and values, stressing that governments must address climate change and sustainable development in parallel, or risk exacerbating poverty and inequality.

Without aggressive action, the world could become an almost impossible place for most people to live in.

The IPCC report includes recent research suggesting that the amount of carbon that humanity can emit while limiting warming to 1.5 °C might be larger than was thought. The previous IPCC assessment, released in 2014, estimated that the world would breach 1.5 °C by the early 2020s at the current rate of emissions. The latest report extends that timeline to 2030 or 2040, on the basis of studies that revised the estimate of warming that has already occurred (R. J. Millar *et al. Nature Geosci.* **10**, 741–747; 2017).

"Every extra tonne of carbon that we dump into the atmosphere today is a tonne that will have to be scrubbed out at the end of the century," says Myles Allen, a climate scientist at the University of Oxford, UK, and one of the lead authors of the report.

"I think we need to start a debate about who is going to pay for it, and whether it's right for the fossil-fuel industry and its customers to be enjoying the benefits today and expecting the next generation to pay for cleaning it up," Allen says.

But scientists have only "medium confidence" in the revised carbon budgets, says Thomas Stocker, a climate scientist at the University of Bern. He says that researchers will provide a more comprehensive look at the numbers in the next full climate assessment, which is scheduled to be released in 2021.

In the meantime, the newer and larger carbon budget could send the wrong message to policymakers, says Oliver Geden, a social scientist and visiting fellow at the Max Planck Institute for Meteorology in Hamburg, Germany. He fears that the IPCC report undersells the difficulty of achieving the 1.5 °C goal. "It's always five minutes to midnight, and that is highly problematic," he says. "Policymakers get used to it, and they think there's always a way out." ■

BIOLOGY

Peer-reviewed homeopathy study sparks uproar in Italy

Homeopathy advocates have championed the paper, but scientists doubt its claims.

BY GIORGIA GUGLIELMI

A study[1] that claims to show that a homeopathic treatment can ease pain in rats has caused uproar after it was published in a peer-reviewed journal. Groups that promote homeopathy in Italy, where there is currently a debate about how to label homeopathic remedies, have held the study up as evidence that the practice works. But several researchers have cast doubt on its claims.

The authors acknowledge some errors flagged in an analysis of the paper by a separate researcher, but stand by their overall conclusions. One of the authors, pharmacologist Chandragouda Patil of the R. C. Patel Institute of Pharmaceutical Education and Research in Dhule, India, also says that the results are preliminary and cannot yet be applied to people, and that he hopes that the team's findings will encourage other researchers to conduct clinical studies.

Researchers have presented evidence in support of homeopathy before — famously, in a 1988 *Nature* paper[2] by French immunologist Jacques Benveniste that was later discredited. This latest claim has attracted attention, in part, because it passed peer review at the journal *Scientific Reports*. (*Nature*'s news team is editorially independent of its publisher Springer Nature, which also publishes *Scientific Reports*.)

"It's worrying that a major journal like *Scientific Reports* didn't pay close attention to a study that claims to show that homeopathy works," says Enrico Bucci, the researcher who carried out the analysis of the paper. Bucci is co-founder of the company Resis in Turin, Italy, which provides tools to uncover potential problems with scholarly articles, and a researcher in systems biology at Temple University in Philadelphia, Pennsylvania.

A paper that claims something as exceptional as the corroboration of homeopathy but also contains errors "raises questions on whether the review process was adequate", adds Michelangelo Cordenonsi, a cancer researcher at the University of Padova in Italy.

A spokesperson for *Scientific Reports*, which published the paper on 10 September, says that the editors are looking into the criticisms, and will correct or retract the paper if necessary. On 1 October, the journal added an editors' note to the homeopathy paper alerting readers to criticisms regarding the study.

HEALING RESPONSE

Homeopathy is based on the idea that illnesses can be treated using substances that produce similar symptoms. Mostly, the substances have been heavily diluted in water or alcohol so that none or only a few molecules of the active ingredient are present. Some supporters of the practice say that the water or alcohol 'remembers' the substance, which triggers a healing response. But these claims aren't backed up by scientific evidence, and the European Academies' Science Advisory Council notes that homeopathic products are no more effective than placebos in treating health problems. ▶

Patil and colleagues report that a homeopathic product (a heavily diluted extract from *Toxicodendron pubescens*, a plant known as Atlantic poison oak) is as effective as the prescription drug gabapentin in reducing inflammation and pain responses in lab-grown cells and animals.

Homeopathy groups worldwide have welcomed the study. And in Italy, where a proposal to label homeopathic products as 'preparations' rather than 'drugs' has provoked heated debate, homeopaths and their associations have said that the study's publication demonstrates the effectiveness of homeopathy.

On social media and in the press, scientists in Italy have voiced concerns about the study. In his analysis, Bucci used his company's software to detect two identical images that supposedly describe different experiments in one of the paper's figures. He also found in the body of the text that the authors write that they had treated the animals with heavily diluted *Toxicodendron pubescens* (up to 1×10^{-30}), but the data in one of the figures show the effects for dilutions up to 1×10^{-8}. These discrepancies, as well as the image duplications, were also flagged by others on PubPeer, a platform to discuss scholarly articles. In another figure, Bucci spotted what seem to be the same data for two different experiments. He published his analysis online on 26 September and sent a detailed report to the editors of *Scientific Reports* on 3 October.

Patil attributes the duplicated images and the repeated data to mistakes that his team made while preparing the manuscript. The discrepancies between the text and the figures are the result of typos, according to Patil. The group will ask *Scientific Reports* to update the article with a correction. But "this does not change the scientific conclusions in any way", Patil says. All the experiments were done "with utmost integrity". The aim of the study was neither to criticize nor to support homeopathy, but to evaluate a homeopathic product using "pharmacological principles", he adds.

Bucci says that he has also found that some of the study's authors, including Patil, had written another paper[3] published in *Scientific Reports* in 2016 that he says also contains inappropriate image duplications. Patil says that these occurred while converting the figures to high resolution when the researchers submitted the manuscript to the journal. The group will ask *Scientific Reports* to correct that article too, he says.

The spokesperson for *Scientific Reports* says that the editors are looking into the issues raised for both papers. "We take our responsibility to maintain the accuracy of the scientific record very seriously." ■

1. Magar, S. *et al. Sci. Rep.* **8**, 13562 (2018).
2. Davenas, E. *et al. Nature* **333**, 816–818 (1988).
3. Chanchal, S. K. *et al. Sci. Rep.* **6**, 30007 (2016).

PUBLISHING

Europe's open-access plan seeks US support

Plan S architect heads to the White House.

BY HOLLY ELSE

A month after European funders launched the 'Plan S' initiative, which demands open access to scientific papers immediately after publication by 2020, the plan's creators have revealed more details — and are seeking support from US policymakers.

"We cannot afford to stand still or slow down. By the end of the year, if we don't have more funders and statements of support, we will miss the boat," says Robert-Jan Smits, the European Commission's senior adviser on open access. Smits was in the United States last week to talk to research funders, scientific societies and representatives of the White House's Office of Science and Technology Policy. "I'm going for business, not chitchat," he told *Nature*.

Smits has also named John-Arne Røttingen, head of the Research Council of Norway, and David Sweeney, executive chair of the funding body Research England, as the leaders of a task force that will decide how funders will implement Plan S. Sweeney accompanied Smits to the United States, along with Marc Schiltz, president of the Brussels-based advocacy group Science Europe, which published Plan S on 4 September.

> ***"We cannot afford to stand still or slow down."***

The task force will release more details by the end of this year, and will consider whether publishers might develop new business models that "outsmart" the plan's requirements, Smits says.

Initially, a coalition of 11 national research funders, including agencies in France, the Netherlands and the United Kingdom, backed the plan; on 24 September, the Academy of Finland joined the group. Plan S funders say that, from 2020, they will require scientists who receive grants from them to make the resulting papers free to read immediately on publication, with a liberal publishing licence allowing others to download, translate or otherwise reuse the work. By contrast, the US National Institutes of Health allows up to one year before papers must be made openly available.

The plan, which aims to flip journals to fully open-access publishing, also states that scientists can't publish in 'hybrid' journals, which collect subscriptions but permit some papers to be published openly for a fee. As written, Plan S would bar researchers from publishing in 85% of journals, including *Nature* and *Science* — unless the journals adapt their business models to open-access publishing. (*Nature*'s news team is editorially independent of its publisher, Springer Nature.)

But details remain unclear. Since the plan's launch, for instance, researchers have wondered whether they would be complying with its intentions if they immediately made a copy of their accepted paper available online — even if the publisher kept the work paywalled.

In mid-September, Smits suggested at the conference of the Open Access Scholarly Publishers Association in Vienna that this would be consistent with Plan S, as long as the open version used a liberal publishing licence. That might mean that paywalled journals could respect Plan S without changing their publishing models. But it is not clear whether this would apply to hybrid journals. Details such as which licence would be acceptable for the archived manuscript, and whether the publisher or the author would retain copyright, also remain fuzzy.

ACADEMIC FREEDOM

Since the plan's launch, an argument has also flared up over whether funders should be able to restrict where academics can publish. Britt Holbrook, a philosopher at the New Jersey Institute of Technology in Newark, co-wrote a blogpost arguing that the plan is unethical because regulating where researchers can publish impinges on academic freedom. His co-authors include some European scientists, such as biochemist Lynn Kamerlin at Uppsala University in Sweden.

But other researchers disagree. Peter Suber, director of the Harvard Open Access Project in Cambridge, Massachusetts, says it is reasonable for funders to restrict how their money is used. Suber says that taxpayer-funded research agencies have a duty to spend their money in the public interest.

Smits says it is a "pity" that the academic-freedom argument is being used, "because it stifles a lot of debate". He says it will be key for the task force to think through the potential consequences of Plan S, to mitigate the risk of publishers developing new business models that the coalition "will regret after five years". "We need to think three moves ahead, like a chess game," he says. ■

BARRY MANSELL/*NATURE* PICTURE LIBRARY

The dusky gopher frog is at the centre of a pitched legal battle over the Endangered Species Act.

LAW

Science and the Supreme Court

The research-related cases awaiting the top US judges.

BY SARA REARDON

The ideological balance of the US Supreme Court shifted to the right on 6 October, as the Senate voted to confirm federal judge Brett Kavanaugh for a seat on the nation's highest court.

Kavanaugh becomes the fifth conservative justice on the nine-member court, which began its latest term on 1 October. *Nature* looks at the science-related cases that are on the court's docket, and others that are likely to advance in the near future.

ENDANGERED SPECIES

The Supreme Court's first case of the term centres on the dusky gopher frog (*Lithobates sevosus*). Development projects have destroyed the amphibian's natural habitat in the southeastern United States, and fewer than 100 of the frogs remain, in a trio of ponds in Mississippi.

To save the species, the US Fish and Wildlife Service (FWS) wants to restore ponds on 2,621 hectares of land in Louisiana owned by timber companies, and then move the animals there. "If you don't do that, the frog is doomed," says Patrick Parenteau, an environmental lawyer at the Vermont Law School in South Royalton.

But the timber companies argue that the pond plan oversteps the bounds of the Endangered Species Act (ESA) of 1973. The law requires the government to protect endangered species' habitats, but does not specify whether this applies to land not currently suitable for a species to occupy.

The case, which justices heard on 1 October, is only the fifth challenge to the ESA to come before the Supreme Court. A ruling in favour of the FWS could pave the way for the government to seize private land and create habitat for other endangered animals and plants — at a time when climate change is rendering many species' longtime habitats unsuitable.

DEATH PENALTY

People whose mental disabilities prevent them from understanding their crime or guilt cannot legally be put to death in the United States. The Supreme Court is poised to decide whether this ban applies to people who were mentally capable when they committed a crime but later developed cognitive impairments.

On 2 October, the court heard arguments in the case of Vernon Madison, who was sentenced to death for murdering an Alabama police officer in 1985. Madison suffered several strokes on death row and is now unable to remember committing the crime. His lawyers say that executing Madison would constitute cruel and unusual punishment.

The state of Alabama argues that Madison can understand its reasoning for putting him to death if the situation is explained to him. But experts say that this understanding is limited. "Madison can mouth the words, but it really comes down to a value judgement: if this person displays these symptoms, is that someone who can prepare themselves for death?" asks Daniel Volchok, an attorney at the firm WilmerHale in Washington DC.

The American Psychological Association and the American Psychiatric Association filed a joint brief in support of Madison. They say that brain imaging and cognitive tests prove that Madison, who is unable to walk or care for himself, cannot truly comprehend the rationale behind his punishment.

CLIMATE CHANGE

The Supreme Court has not yet agreed to hear any cases this term involving climate change. But the Trump administration has worked to roll back a wide range of climate regulations, prompting a wave of lawsuits. Some of those will reach the Supreme Court, says Sharon Jacobs, an environmental lawyer at the University of Colorado Boulder.

Potential cases this term include a challenge to the Federal Energy Regulatory Commission's decision to limit consideration of climate change when it evaluates applications for new natural-gas pipelines. Earlier this year, the agency said it would no longer require companies to address the climate impact of burning the gas in their licence applications.

Another suit that could end up on the docket seeks to limit the reach of the Clean Air Act. The law, which took effect in 1990, banned chemicals called chlorofluorocarbons (CFCs) that destroy the ozone layer. Some manufacturers then switched to using hydrofluorocarbons (HFCs), which do not deplete ozone but are powerful greenhouse gases.

In 2015, the US Environmental Protection Agency (EPA) ordered two companies, Mexichem Fluor and Arkema, to switch to less-harmful chemicals — and the firms fought back, arguing that the Clean Air Act applies only to CFCs. Last year, Kavanaugh wrote a lower-court ruling that said the EPA could not require companies that were using HFCs to replace them with less-damaging chemicals. Environmental groups and companies that make replacements for HFCs appealed on the agency's behalf.

If the case reaches the Supreme Court, Kavanaugh will probably recuse himself — increasing the likelihood of a split decision. ■

CHEMISTRY

Nobel for test-tube evolution

Controlling protein evolution in the lab has led to greener technologies and new medicines.

BY ELIZABETH GIBNEY, RICHARD VAN NOORDEN, HEIDI LEDFORD, DAVIDE CASTELVECCHI & MATTHEW WARREN

Ways to speed up and control the evolution of proteins to produce greener technologies and new medicines have won three scientists the 2018 Nobel Prize in Chemistry.

Chemical engineer Frances Arnold, at the California Institute of Technology in Pasadena, is just the second woman to have won the prize in the past 50 years. She was awarded half of the 9-million-Swedish-krona (US$1-million) pot. The remaining half was shared between Gregory Winter at the MRC Laboratory of Molecular Biology in Cambridge, UK, and George Smith at the University of Missouri in Columbia.

Arnold carried out pioneering work in the 1990s on the 'directed evolution' of enzymes — proteins that catalyse chemical reactions. She devised a method for inducing mutations in enzyme-producing bacteria and then screening and selecting the bacteria to speed up and direct enzyme evolution. These enzymes are now used in the production of biofuels and drugs.

"Biology has this one process that's responsible for all this glorious complexity we see in nature," she told *Nature* shortly after the prize announcement on 3 October. But whereas nature operates blindly, Arnold's techniques accelerate natural selection towards producing enzymes with known properties. "It's like breeding a racehorse," she says.

In 1985, Smith pioneered a technique that uses a bacteriophage — a virus that infects bacteria — as a host that displays a protein on its outer coat, allowing researchers to find other molecules that interact with the protein. Winter developed and improved this technology, called phage display, and invented ways to use it to evolve antibodies adapted for use as therapeutics. Today, antibodies evolved using this method can neutralize toxins and counteract autoimmune diseases.

Nobel laureates Gregory Winter (left), Frances Arnold and George Smith.

L–R: AGA MACHAJ; CALTECH; UNIV. MISSOURI-COLUMBIA

The first humanized antibody, called adalimumab (Humira), was discovered by Cambridge Antibody Technology — a company that Winter co-founded in 1989 — and was approved for treating rheumatoid arthritis in 2002. It is also used to treat psoriasis and inflammatory bowel diseases. In 2017, it was the world's top-selling drug, generating revenues of $18.4 billion.

Scepticism abounded when the company was launched, says co-founder David Chiswell, and it struggled to find investors. "Nobody in the world believed that antibodies were really good," says Chiswell, who is now chief executive of Kymab, an antibody company in Cambridge.

Arnold also faced a battle when she put forward the idea of evolving proteins in the lab, says Dane Wittrup, a protein engineer at the Massachusetts Institute of Technology in Cambridge. Researchers thought then that they would be able to sit down at a computer and rationally design proteins to carry out specific functions. "But now, by and large, directed evolution is how the work is done."

Winter says that a woman with cancer who had received an early, experimental version of one of his humanized antibodies against a cancer-related protein drove him to push his research out of the laboratory and into the clinic. When Winter warned her that the effects of the therapy might not last, she told him she only needed to live for a few more months, so that she could help her dying husband. "I was so choked by that," Winter says.

Before Arnold, the last woman to win the Nobel Prize in Chemistry was Ada Yonath, a crystallographer at the Weizmann Institute of Science in Rehovot, Israel, who won in 2009 for mapping the structure of the ribosome, which generates proteins from the genetic code in cells. Before her, the most recent woman to win was crystallographer Dorothy Hodgkin, in 1964. Arnold is just the fifth female winner in the prize's history. ■

EUROPE

Tiny space fleet could track CO_2

Project could help to show whether nations are meeting pledges to cut emissions.

BY ALEXANDRA WITZE

European researchers are developing a miniaturized instrument that could precisely measure carbon dioxide coming from cities and power plants. If it works, the device could fly aboard a constellation of small satellites starting in the late 2020s, helping to track daily fluctuations in greenhouse-gas emissions.

Developers with the 3-year, €3-million (US$3.5-million) project envisage it complementing more-expansive efforts to monitor CO_2 from space, such as a proposed set of new Sentinel Earth-observing satellites from the European Space Agency. If approved, those might also come online in the late 2020s.

Several satellites currently monitor CO_2 emissions, including Japan's GOSAT, the United States's Orbiting Carbon Observatory-2 (OCO-2) and China's TanSat. But none of them launched with the explicit goal of tracking compliance with global treaties.

In 2015, before the signing of the Paris

accord to limit greenhouse-gas emissions, the European Commission began exploring how it could develop satellites to assess whether nations are abiding by their climate pledges.

The new, small sensor could play a part in that. “We want to improve the accuracy of monitoring anthropogenic CO_2 emissions,” says Laure Brooker Lizon-Tati, an engineer with Airbus Defence and Space in Toulouse, France. She coordinates the project, called the Space Carbon Observatory (SCARBO), which is being developed by a consortium of eight European companies and research institutions.

Team scientists were scheduled to describe the first results at a space-optics conference this week in Chania, Greece.

The proposed Sentinel satellites would precisely measure greenhouse gases around the world. But they would not be able to make daily measurements above places of interest, such as cities. “This is where a constellation of tiny SCARBO systems could come into the game,” says Heinrich Bovensmann, a remote-sensing researcher at the University of Bremen in Germany.

SMALL STEPS

SCARBO satellites would weigh just 50 kilograms each, roughly one-tenth the mass of OCO-2 or TanSat. An estimated two dozen working together would be able to cover the globe once a week, and could fly over particular areas of interest once a day. Together, they could monitor frequent changes in carbon emissions, such as morning and afternoon surges from an industrial area.

But first, SCARBO scientists have to show that their plan can work. At its heart is a miniaturized spectrometer — no longer than an outstretched hand — that would detect CO_2 concentrations in the air below.

“We want to improve the accuracy of monitoring anthropogenic CO_2 emissions.”

Fitting a spectrometer onto a small satellite requires shrinking optics and developing new methods for analysing CO_2 concentrations. “It’s a real challenge,” says Bovensmann.

The scientists’ goal is to measure CO_2 concentrations to an accuracy of less than 1 part per million at a resolution of 2 kilometres — comparable to the data collected by larger satellites now in orbit.

“We want to prove the technology can achieve these types of measurements,” says Etienne Le Coarer of the University Grenoble-Alpes in France, which is building the instrument along with the ONERA French aerospace laboratory in Palaiseau.

NASA’s Jet Propulsion Laboratory in Pasadena, California, has worked on a similar concept for miniaturized sensors, but using a different type of spectrometer.

SCARBO scientists plan to test their instrument aboard a research aeroplane in 2020. It will fly alongside a Dutch-built instrument to study atmospheric aerosols, which are a major source of error when trying to measure greenhouse gases. The test will be the first time that aerosols and CO_2 are measured simultaneously to improve the quality of data on greenhouse-gas emissions, says Lizon-Tati.

SCARBO is focusing on CO_2 monitoring, although it would also be useful for tracking methane emissions, says Le Coarer. Several private efforts to monitor methane emissions cheaply from space are already under way, including a Canadian microsatellite that has been flying since 2016 and a planned small satellite from the Environmental Defense Fund, an advocacy group in New York City. ■

CORRECTION

The News story ‘Peru plans oil clean-up’ (*Nature* **562**, 18–19; 2018) erroneously stated that the United Nations Development Programme (UNDP) funded a study on remediation strategies. In fact, the Peruvian government funded the study and the UNDP conducted it.

FORTRESSES OF MUD

Rising seas threaten the San Francisco Bay Area, home to one of the largest estuaries in North America. But marsh-restoration efforts could hold back the high water.

BY ERICA GIES

There's something apocalyptic about this pond on the east side of San Francisco Bay, California. The legacy of a salt industry that has moved elsewhere, it has subsided a couple of metres below the level of neighbouring marshland. Algae paints red swirls in the brown water, and the pond's edge is crusted hard with sparkling salt. As a breeze eases off the bay, a squadron of pelicans sails by, en route to more-appetizing hunting grounds.

But there is a better future ahead for landscapes like this one in the Eden Landing Ecological Reserve and elsewhere around the bay. Over the next decade, government officials plan to fill many such depressions with sediment and then open them up to the tides. Eventually, cordgrass, pickleweed and other marsh vegetation will take root, restoring this crucial marsh ecosystem. The goal is to try to create a natural buffer to protect the heavily populated waterfront, by sapping energy from storm surges and blocking the highest tides.

San Francisco Bay's salt ponds are part of a much broader story. After a century of human development destroyed most of the area's wetlands, the region did an about-face in the 1970s. It became a leader in marsh restoration, moving into high gear after a groundbreaking plan published in 1998. In recent years, local leaders have tackled these efforts with a new-found sense of urgency. Sea levels here could rise by as much as 2.1 metres by 2100, the California Natural Resources Agency estimates, and that would threaten electricity plants, transportation infrastructure and drinking-water facilities in the region — many of which lie low and close to the bay.

A view of a wetlands restoration project in Menlo Park, California.

MICHAEL MACOR/*SAN FRANCISCO CHRONICLE*/POLARIS/EYEVINE

Marshes have a superpower in the fight against sea-level rise. Unlike artificial barriers such as sea walls and levees, they can evolve, growing progressively higher as they trap more sediment and their vegetation decomposes and regrows. "Marshes are in a dynamic equilibrium with the water level," says John Bourgeois, executive manager of the South Bay Salt Pond Restoration Project, a public–private partnership that manages wetlands restoration in Eden Landing and other sites in the South Bay. "It's been clearly shown that, even at pretty high rates of sea-level rise, if there's enough suspended sediment, they can keep pace."

San Francisco Bay is not the only region where ecosystems are being enlisted in the fight against climate change. Around the world, researchers and governments are looking to natural coastal infrastructure, including dunes, gravel beaches and mangroves, to protect communities from flooding. San Francisco Bay's efforts are among the oldest, having already restored to the tides about 8,000 hectares of habitat. Techniques developed by local scientists have been adopted by US agencies and applied elsewhere in the world, says Peter Goodwin, president of the University of Maryland Center for Environmental Science in Cambridge, Maryland.

But the Bay Area faces challenges in constructing a protective phalanx of marshes. Aside from the cost and the outdated regulations that

slow work, one of the biggest hurdles is finding enough sediment to do the job. Human development has trapped sediment behind dams and levees, leaving the bay, like many deltas around the world, without enough raw material to keep up with rising seas. Researchers will need to locate large quantities of sediment to fill in sunken former marshes and jump-start the restoration process. And they expect that they will also need to deliver sediment to existing marshes as sea levels rise.

Other researchers are watching to see how the Bay Area's experiments — and its close partnerships among government agencies, scientists and others — pan out. Every coast may be unique, says Eugene Turner, a professor of oceanography and coastal sciences at Louisiana State University in Baton Rouge, but they all face common issues, such as rising seas, changing temperatures and "inappropriate development along the ocean's edge", so projects can learn from each other.

RETURN TO NATURE

On a summer day on the northern edge of San Francisco Bay, a marsh harrier glides above a low-lying expanse of tawny yellow and variegated green vegetation, cut with curvaceous channels. Such sights were not as common decades ago. By the 1950s, all but 8% of the bay's 77,000 hectares of marshes had been dyked off or filled for human uses such as agriculture, rubbish dumps, sewage-treatment plants, navy bases, airports and salt ponds.

Growing public awareness of that devastation turned to outrage in the 1960s. In response, local scientists began some of the world's first marsh restoration projects. Since then, hundreds of researchers, policymakers and regulators have worked together to set and meet ambitious goals. As of 2015, marshlands and mudflats that are in the process of becoming marshes occupied about 29% of the historic marsh area. And local groups have purchased another 10% or so — about 7,300 hectares — and slated it for restoration (see 'More marshes').

In the beginning, restorers were motivated by a desire to recreate habitats for endangered species, prevent flooding and provide recreation areas. But in the past two years, protecting the region from sea-level rise has become an explicit objective. In a 2016 report (see go.nature.com/2djjojb), locals set a new goal to restore as much marsh acreage as possible by 2030, in the hope that the marshland will be high enough and strong enough to keep up with the expected acceleration in sea-level rise by the middle of the century.

Turning to natural systems for such protection is a radical shift from the concrete engineering that dominated the last century. But as human populations expand and climate impacts intensify, the costs and limitations of the artificial approach are becoming more apparent. Sea walls, for example, are brittle; they break rather than flex and have unwanted effects on surrounding areas by, for instance, diverting wave energy to unprotected locations. Coastal ecosystems not only protect inland areas but also provide habitat for endangered species, fish nurseries and natural water-cleaning services. In addition, they can help to combat climate change by storing carbon as they grow.

The United States is conducting some of the world's largest wetlands restoration efforts. To staunch its dramatic land loss, for example, Louisiana is planning multiple diversions to siphon sediment from the Mississippi River to replenish wetlands. The states that surround the Chesapeake Bay on the US east coast aim to restore more than 40,000 hectares of wetlands — although the primary motivation is to reduce pollution from agriculture and urban runoff. In Europe, most tidal-marsh restoration projects, like an effort to build a marsh near the Dutch port of Delfzijl, are much smaller in scale.

Other regions, such as Long Island in New York, and southern California, as well as countries including England, Singapore and China, regularly enquire about the Bay Area's projects and methods, says Letitia Grenier, director of the resilient landscapes programme at the San Francisco Estuary Institute (SFEI), an independent aquatic- and ecosystem-science institute based in Richmond, California. "The science here is very proactive and very progressive in the sense of using natural processes as the solution," she says. Researchers outside the Bay Area note that a key strength is its process. "It's probably one of the best examples in the country of how universities, agencies and NGOs are working together — and the private sector more recently," says Goodwin.

But despite widespread regional support for wetlands restoration, big obstacles remain to building enough marshland to protect against sea-level rise. Recreating tidal marshes requires three ingredients: time, space and sediment. Bourgeois says it can take between 5 and 20 years before a salt pond fills with enough sediment for a marsh to take root. Space is also at a premium. Marshes grow upwards in part by marching slowly inland. Unfortunately, says Grenier, "what we have in the back of our marshes is freeways and sewage-treatment plants, and Oracle and Google".

SEDIMENTARY EFFORTS

What's more, there's a major sediment deficit. Salt ponds such as those at Eden Landing have subsided so far below the level of surrounding marshland that, in order to speed up their recovery, restorers will need to fill them with extra sediment from elsewhere before opening them to the tide. Researchers think that it will also be necessary to feed existing marshes to help them keep pace with rising sea levels. Scott Dusterhoff, the SFEI's lead geomorphologist, estimates that existing marshes, along with the oyster-bearing mudflats between them and the bay, will face a deficit of roughly 100 million metric tonnes of sediment by 2100 if sea levels rise by 1 metre, a middle-of-the-road scenario. If current trends continue, scientists fear that most of the bay's marshes will be damaged or destroyed by 2100.

Natural marshes receive sediment from two directions: upstream rivers and ocean tides. But in San Francisco Bay, as in many river deltas around the world, these systems are disconnected. The sediment supply is blocked by dams across rivers and levees against tides. Because so much sediment is needed to fill the subsided areas, restorers will have to get creative about sourcing sediment, says Bourgeois. "Everything is on the table at this point."

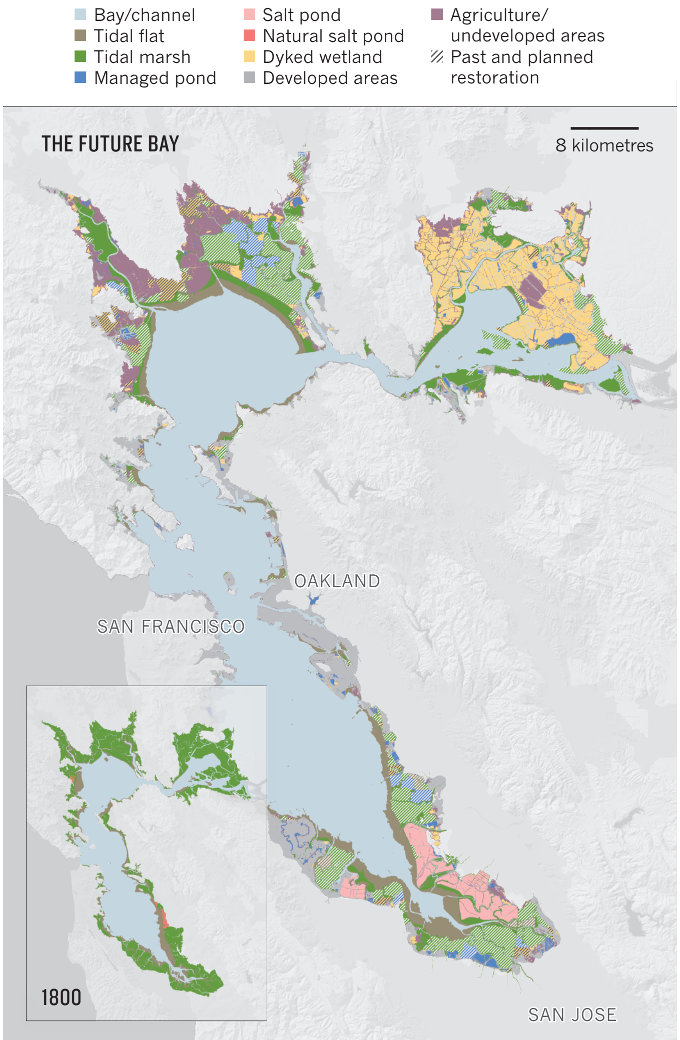
One large potential source lies in San Francisco Bay's deepwater ports, which require routine dredging to remove sediment piled up by the tides. Much of that dredged material is carried offshore and deposited in the ocean, says Brenda Goeden, sediment programme manager for the San Francisco Bay Conservation and Development Commission, a state planning and regulatory agency. The practice was begun to protect the bay from dumped sediment, which can harm wildlife if it is contaminated or clouds the water. Eventually, in the 1990s, regulatory agencies recognized the need for sediment in marsh restoration and began issuing permits for the 'beneficial use' of dredged material that was proved to be clean. But last

MICHAEL LAYEFSKY

Salt ponds dominate this aerial view in Redwood City, California.

MORE MARSHES

By the mid-twentieth century, human activity had eliminated more than 90% of the tidal marshland that once rimmed California's San Francisco Bay. Restoration efforts aim to bring the tidal marsh area back to nearly 40% of its historic reach.



SOURCE: SFEI

year, about half was still deposited in the ocean, says Goeden.

Dirt could also come from the region's construction boom. SediMatch, a project by the non-governmental organization San Francisco Bay Joint Venture and the SFEI, is putting "folks with sediment together with people doing bay restoration who need sediment", says Grenier. "It's a dating service for sediment." But the pace of moving material can be slow. The salt ponds at Eden Landing, for example, require an estimated 5.4 million cubic metres of sediment to bring them close to the level of surrounding marshland. A truck delivers about 8.4 cubic metres at a time. "Do the math," says Bourgeois. "That's a lot of trucks." He is proposing an alternative: a barge in the middle of the bay where dredged sediment could be mixed with water to create slurry, which could then be delivered by pipe to Eden Landing and other restoration projects.

But questions remain about how best to deliver scavenged sediment, both to lift defunct marshes quickly and, in the future, to help existing marshes to keep pace with sea-level rise.

There is a range of possible tactics. In Louisiana, for example, a project called Bayou Dupont is dredging Mississippi River mud and moving it through a kilometres-long pipeline to build wetlands. The marshes there have been so severely starved of sediment — and affected by subsidence, erosion, and oil and gas development — that they are drowning rapidly. So the project sprays dredged sediment mixed with water directly down onto the sinking marsh to raise its elevation quickly, as Bourgeois is proposing for some of the South Bay sites that are not yet active marshes.

But raining slurry "fills in all the holes and channels simultaneously, so it tends to flatten out the topography", says Jeremy Lowe, a senior environmental scientist at the SFEI. Marsh plants and animals prefer more variability, and varied topography is better for slowing down floodwaters, too. For these reasons, restoration ecologists in the Bay Area typically stop filling about 30 centimetres below the marsh plain level, then open the area to tidal and possibly river flows, allowing natural systems to finish the job.

The US Army Corps of Engineers and the SFEI want to test the use of waves and tides to move supplemental sediment into place, not only to avoid the uniformity problems created by slurrying, but also to avoid its cost and energy footprint. This approach could become necessary if future marshes need supplemental sediment to keep up with sea-level rise; the material will need to be delivered more delicately to a functioning wetland than to a salt-pond hole. The Bay Area is angling to take a "kinder, gentler, more natural-process approach", says Grenier, that will allow "plant and animal populations to thrive".

As part of that as-yet-unfunded effort, the groups also hope to study how much deployed sediment actually lands on marshes targeted for restoration, which would help scientists to understand how much sediment is needed and how long restoration will take. Lowe is proposing to use fluorescent tracers to track where particles go in the bay. Another experiment with a soft approach to sediment deposition is under way in the Netherlands. The pilot project takes sediment dredged from the port at Harlingen on the Wadden Sea, and releases it farther along the coast, where researchers hope it will help to shore up a tidal marsh at Koehoal.

Additional river sediment could also be delivered to Bay Area marshes by freeing up blockages behind dams and by slowing material that flows quickly through flood-control channels, causing it to shoot past marshes and into deeper water. In some cases, existing dams could be operated differently to mimic natural, periodic pulses that flush sediment. Newer dams can spill water from the bottom, says Lowe, allowing sediment to pass. Restoring upstream sediment supplies could also help to protect the chain of coastal habitats — including uplands, high and low marshes, mudflats and subtidal zones — that, when free from human-made barriers, exchange critical nutrients and materials. "If you just have the marsh, it's not as resilient as if you have the full system," says Grenier, "because each element protects what's behind it."

MARSH MOMENTUM

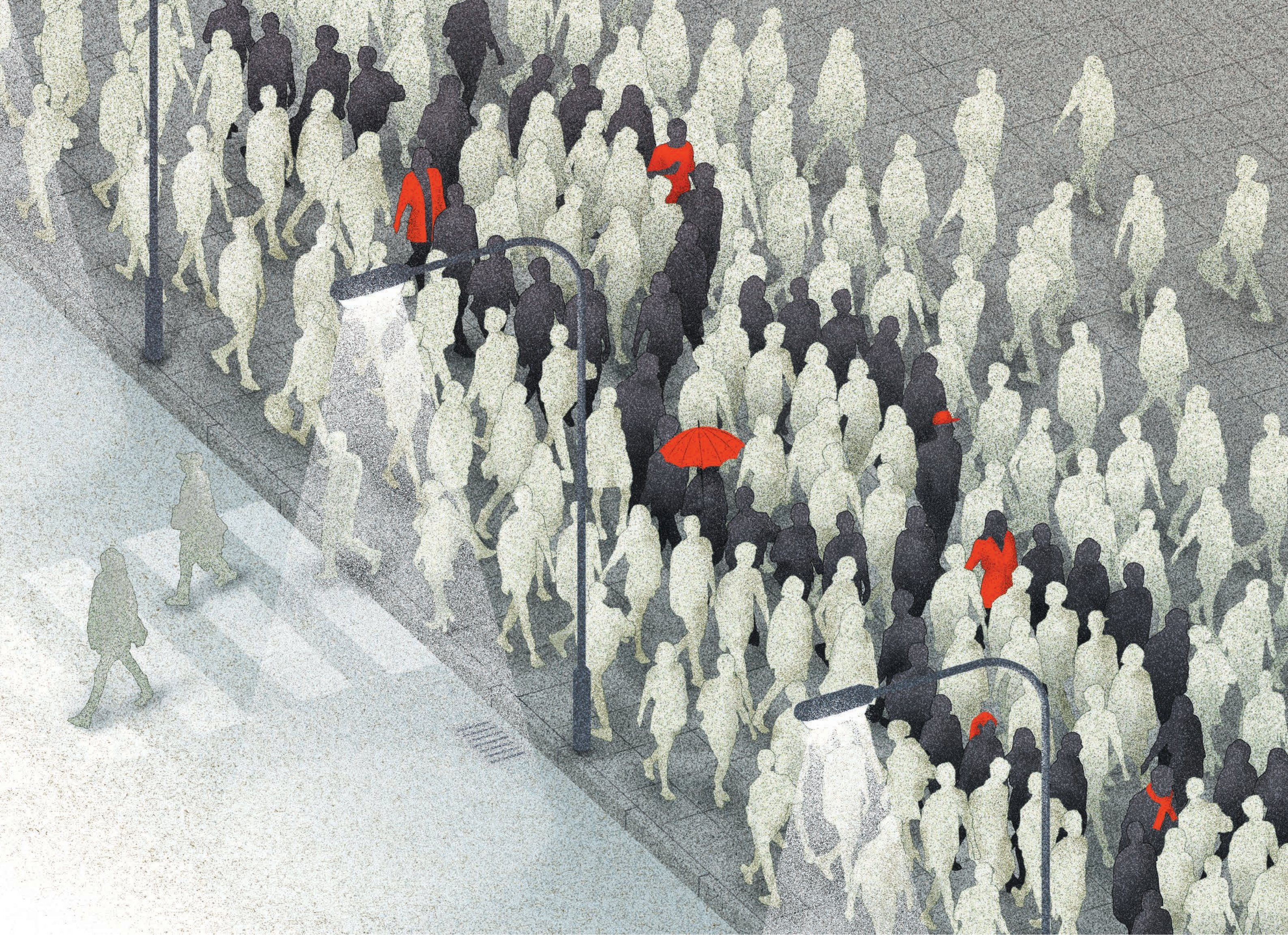
San Francisco Bay's tidal-marsh restoration efforts are gaining momentum. In 2016, voters in the nine Bay Area counties overwhelmingly approved Measure AA, a tax that is expected to raise US$500 million over 20 years for marsh restoration and related flood-control projects. "We're showing that the region is extremely serious about this," says Goeden. "We're willing to put our money where our mouth is."

That money won't be enough to do the job, but local advocates are hoping that they can leverage it to get federal funding. The clock is ticking for the restoration community to meet its 2030 goal. If it can find sediment to fill subsided areas, marshes could begin to establish themselves in 1–5 years, depending on the starting elevation and other variables, says Bourgeois.

The agencies responsible for dredging regulations recognize the threat of sea-level rise and are currently re-evaluating their policies, but regulatory hurdles remain. "A lot of what we're proposing presents a real struggle to get permits because all of the laws were focused on preventing people from filling the bay," says Bourgeois. He and other restorers say permits aren't being issued fast enough to meet their 2030 restoration goal.

In the end, all of this is one big experiment. "Restoring marshes is still somewhat new to society, and so there will be surprises, mistakes and unexpected successes," says Turner. The Bay Area's effort is doing many positive things, he says: storing carbon underground, recreating wildlife habitat and investing in natural infrastructure, which is much cheaper than emergency relief after disasters. Yet, he adds, "It is unclear if any coastal restoration programme will be enough to resist sea-level rise". ■

Erica Gies *is an independent journalist based in Victoria in Canada, and San Francisco.*


NIK SPENCER/NATURE

The power of many

Health predictions based on the make-up of the human genome have taken a great leap forward. But polygenic risk scores are still highly controversial.

BY MATTHEW WARREN

6.6 million — that's how many spots on the human genome Sekar Kathiresan looks at to calculate a person's risk of developing coronary artery disease. Kathiresan has found that combinations of single DNA-letter differences from person to person in these select locations could help to predict whether someone will succumb to one of the leading causes of death worldwide. It's anyone's guess what the majority of those As, Cs, Ts and Gs are doing. Nevertheless, Kathiresan says, "you can stratify people into clear trajectories for heart attack, based on something you have fixed from birth".

Kathiresan, a geneticist at Massachusetts General Hospital in Boston, isn't alone in counting outrageously high numbers of variants. The polygenic risk scores he has developed are part of a cutting-edge approach in the hunt for the genetic contributors to common diseases. Over the past two decades, researchers have struggled to account for the heritability of conditions including heart disease, diabetes and schizophrenia. Polygenic scores add together the small — sometimes infinitesimal — contributions of tens to millions of spots on the genome, to create some of the most powerful genetic diagnostics to date.

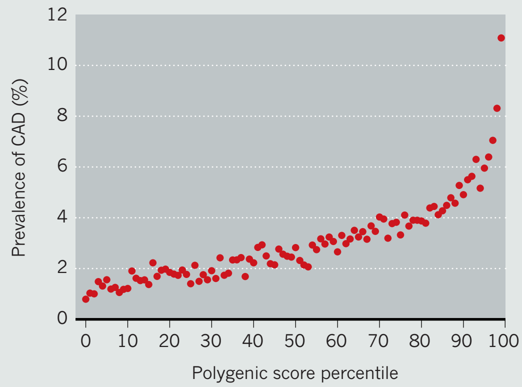
This approach has taken off thanks to a number of well-resourced cohort studies and large data repositories, such as the UK Biobank (see pages 194, 203 and 210), which collect vast quantities of health information alongside DNA data from hundreds of thousands of people. And some studies published in the past year or so have been able to analyse more than a million participants by combining information from such sources, increasing scientists' ability to detect tiny effects.

Supporters say that polygenic scores could be the next great stride in genomic medicine, but the approach has generated considerable debate.

Some research presents ethical quandaries as to how the scores might be used: for example, in predicting academic performance. Critics also worry about how people will interpret the complex and sometimes equivocal information that emerges from the tests. And because leading biobanks lack ethnic and geographic diversity, the current crop of genetic screening tools might have predictive power only for the populations represented in the databases.

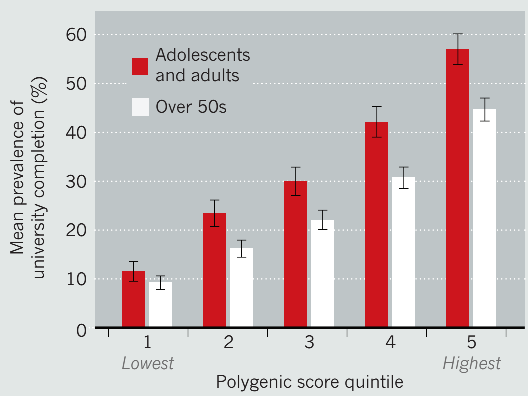
The multi-gene prediction tools

When researchers evaluated polygenic risk scores for coronary artery disease (CAD) in 290,000 people from the UK Biobank, they found that the prevalence of disease rose sharply in the highest percentiles.



SOURCE: REF. 1

Another group tested a polygenic predictor for educational attainment on 5,000 US adults and adolescents and on 9,000 people over the age of 50. Its predictive power was about on par with demographic factors.



SOURCE: REF. 5

"Most people are keen to have a decent debate about this, because it raises all sorts of logistical and social and ethical issues," says Mark McCarthy, a geneticist at the University of Oxford, UK. Even so, polygenic scores are racing to the clinic and are already being offered to consumers by at least one US company.

Peter Visscher, a geneticist at the University of Queensland, Australia, who pioneered the methods that underlie the trend, is broadly optimistic about the approach, but is still surprised by the speed of progress. "I'm absolutely convinced this is going to come sooner than we think," he says.

RISK CALCULATION

When researchers completed the first drafts of the human genome in the early 2000s, many expected that it would mark the start of a medical revolution. Geneticists started searching for the differences that might explain why one person develops diabetes or heart disease whereas another does not. The idea was simple: compare a group of people with the condition to a group without and look for differences in their DNA. The variations generally came in the form of DNA-letter swaps, known as single nucleotide polymorphisms, or SNPs. If people with a condition tended to have a T at a certain location whereas others had a C, that suggested that the SNP was associated in some way with the disease.

These genome-wide association studies — or GWASs, as they came to be known — became very popular. But after years of searching, scientists could still only explain a small bit of the inherited risk for common diseases. It turned out that most of these conditions were related to many more SNPs than scientists had first expected, says Ali Torkamani, a geneticist at the Scripps Research Institute, La Jolla, California.

Worse still, a majority of the variants conferred a very small risk — detectable only when surveying huge groups of people. "We didn't have the sample size to really drive prediction as well as some people naively thought," says Ewan Birney, director of the European Bioinformatics Institute in Hinxton, UK. By 2007, geneticists were fretting about something they called "missing heritability". It was clear that many of these conditions had a genetic component, but GWASs clearly weren't catching much of it.

Today, things are changing. With access to massive data sets, as well as advances in how data are analysed, scientists are getting better at measuring those very small risks, says Kathiresan.

A prime example is the technique Kathiresan used to generate his 6.6-million SNP score, which was published in August[1]. He and his team took data from a 2015 meta-analysis that combined 48 GWASs, consisting of 61,000 people with coronary artery disease and 120,000 controls[2]. They then tested their polygenic predictor on 290,000 people in the UK Biobank, finding that those scoring in the highest few percentiles had on average several times higher risk of developing the disease than did the rest of the population (see 'The multi-gene prediction tools'). Of the 23,000 people who received the highest scores, for example, 7% had coronary artery disease, compared with 2.7% of the remaining population. The group conducted similar analyses for four other disorders, including inflammatory bowel disease and breast cancer, each time identifying a group who scored in the top few percentiles and were at particularly high risk.

The paper has drawn praise from some researchers as a demonstration that polygenic risk scores could, in theory, be used in the clinic. The ability of the scores to identify high-risk groups, Kathiresan says, parallels existing measures of risk used in medicine. "Essentially what you have is a new risk factor for coronary artery disease."

Kathiresan's work made headlines and triggered some controversy — owing to the sheer number of variants included in the risk score. Only a fraction of those 6.6 million SNPs actually contribute to the prediction, says biostatistician Nilanjan Chatterjee from the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, who was not involved in the study. This is because of how these kinds of scores are calculated: data for all the variants are stuck into an algorithm, which assigns a weight to each one according to how strongly it is related to the disease, and most will in fact pose little or negligible risk.

Many researchers, including Chatterjee, say that it doesn't matter if many variants with minimal effect are included. But others worry that including millions of variants that don't do anything could undermine public trust in the scores. Cecile Janssens, an epidemiologist at Emory University in Atlanta, Georgia, says she is not impressed by the study. One of her concerns is that the millions of variants used to calculate the final score didn't improve performance by much compared with a score made from just 74 SNPs with the strongest links to disease. If these sorts of scores are going to be used clinically, she says, "the credibility of the score is also important."

COURSE OF ACTION

Whereas Kathiresan's study focused mainly on genetic risk, others are looking at how the polygenic scores might complement existing measures of risk. In 2013, Samuli Ripatti, a statistical geneticist at the University of Helsinki, found that combining a polygenic risk score with conventional risk factors for coronary artery disease, such as high body-mass index and elevated blood pressure, improved predictions of who would develop the disease[3]. He was also able to identify a group of people with high genetic risk scores who would otherwise have only been considered to be at intermediate risk, and Ripatti says that this ability to pick out individuals who fly under the radar is the biggest benefit of polygenic risk scores.

Genetic risk scores could also improve screening regimes for diseases such as breast cancer. In the United States, women are currently advised to start getting mammograms from the age of 50, but if younger at-risk

women could be identified, they might benefit from earlier screening. In 2016, Chatterjee developed a model for breast cancer that incorporated both conventional risk factors and a polygenic score calculated from around 90 SNPs[4]. On the basis of these scores, he predicted that 16% of women aged 40 have a risk equivalent to the average 50-year-old — suggesting that they could benefit from screenings starting at 40. The team is now testing its model in other data sets and with a larger number of SNPs, to see whether the predictions hold up.

Meanwhile, personalized-medicine company Myriad Genetics in Salt Lake City, Utah, has already begun to include a polygenic risk score for breast cancer in the results it provides to some women. Only about 10% of women with a family history of breast cancer have one of the harmful single-gene mutations associated with the disease, so the company is now returning a score to the remaining 90% that tells them their likelihood of developing breast cancer according to a combination of polygenic risk and factors such as history and lifestyle. One of the strengths of these scores is that they provide a result for everyone, says Jerry Lanchbury, Myriad's chief scientific officer. Although the current focus is on identifying women who are at high risk, in the future he could see the scores being used to find those who are at lower-than-average risk, who might potentially benefit from having less-frequent mammograms. "We start to enter a world where you can provide a precision-medicine result for everyone," Lanchbury says.

ALL IN THE STATISTICS

One complaint about polygenic scores is that they throw out biology in favour of statistics. Polygenic scores alone won't provide much insight for drug development, but the studies can provide a starting point for delving into the individual variants and working out which genes they affect and the mechanisms that might lead to disease.

Part of that insight will come from disentangling which variants actually produce a given trait or disease, and which are just along for the ride. A SNP that is associated with a disease isn't necessarily its cause: it could simply be that the variant tends to be inherited alongside another part of the genome that is directly involved. For example, Kathiresan estimates that only about 6,000 of his 6.6 million SNPs are causally related to coronary artery disease. As sample sizes get larger, it becomes easier to tease these variants apart, says McCarthy.

"It raises all sorts of logistical and social and ethical issues."

There is also still a significant portion of genetic risk that current studies can't account for. Ripatti estimates that 30–50% of the risk for many common diseases is genetic — much of the rest is determined by environmental factors. But the problem of missing heritability remains: as a rule of thumb, GWASs can currently account for about one- to two-thirds of the inherited risk of disease, says Visscher. As sample sizes get larger, researchers will probably find more variants that contribute to the risk, says Torkamani, although the returns diminish. "At some point, you're just going to stop getting too much utility from additional genetic risk factors," he says. More of the genetic risk might also be picked up by whole-genome sequencing, adds Visscher. Currently, GWAS research is conducted mainly using arrays that sequence only a portion of the genome, but as whole-genome sequencing becomes cheaper and more widespread, less-common variants that contribute to disease might become easier to find.

FROM LAB TO CLINIC

Kathiresan says he hopes to have a score for coronary artery disease on the market in the next year. But most researchers acknowledge that there are obstacles to overcome before these scores can be used widely. The number one hurdle, says McCarthy, is applying them to different populations. The risk scores are generated and validated in data sets made up mainly of people with European ancestry, such as the UK Biobank, limiting the extent to which they can be applied to people of other ethnicities. Myriad's score, for example, is currently available only to individuals with a European background, although Lanchbury says that the company is in the process of developing a similar score for African American women. McCarthy says that the ultimate aim is to generate risk scores that are specific to ethnicity.

Ethnicity isn't the only complicating factor, Birney adds. The populations analysed in the studies come from specific health-care systems, and their experiences don't necessarily translate across countries. The chance of having a heart attack could vary between the United Kingdom and United States, for example, as could the standards of care. So scores might not be translatable.

Even the simple act of communicating these scores to people brings with it a number of concerns. Doctors are not necessarily trained in genetics, says McCarthy, and "there aren't enough genetic counsellors on the planet" to conduct the nuanced discussions that genetic risk scores will entail. There is a popular misconception that because our genetics doesn't change, "it's somehow a destiny that will be fulfilled", says Birney. Janssens worries that if people think that the chance of getting a disease is hard-wired into their DNA, they won't be motivated to do anything about it.

The concern becomes even more acute for non-disease traits that might be predicted by such scores. A study on more than 1 million people published earlier this year developed a polygenic score that essentially correlates with how long people stay in education[5]. The authors of that study went to great lengths to clarify they were not suggesting any kind of intervention for people who have extremely low scores. "Any practical response — individual or policy-level — to this or similar research would be extremely premature," they write.

Michelle Meyer, a bioethicist at Geisinger Health System and a co-author on the study says that the score simply isn't actionable. Without understanding the biological differences represented by the score — or the environmental and social factors bound to interact with those differences — it's impossible to know how to intervene.

TALKING GENETICS

Understanding how people will react to polygenic scores is a high priority for researchers. Ripatti and his colleagues have given more than 7,000 individuals in Finland information about their likelihood of developing heart disease, based on both polygenic scores and conventional risk factors such as high blood pressure. Most of the respondents say that getting this information motivates them to make positive changes, says Ripatti. Preliminary results suggest that those with high genetic risk are the most likely to take actions such as losing weight or stopping smoking.

In nearby Estonia, researchers are in the process of genotyping 100,000 individuals, adding to the 50,000 the country has already sampled. And unlike many other biobanks, participants in the Estonian project can sign up to receive feedback. Among the results being returned to them are polygenic risk scores for type 2 diabetes and cardiovascular disease, says Lili Milani, a geneticist at the Estonian Genome Center at the University of Tartu, Estonia. Similar to the Finnish work, participants are shown graphs of how lifestyle changes could reduce or increase their risk. And, says Milani, initial indications are that people are glad for the advice.

For now, people are receiving their scores from genetic counsellors. But Milani is working with the Estonian government to work out how to integrate genomic data into the health-care system, so that it can be used every day by doctors. The country ultimately aims to genotype anyone who's interested, right up to its entire population of 1.3 million, Milani says. "The goal is to build something so great that all doctors will want to recommend it and all of the population will want it." ■

Matthew Warren *writes for* Nature *from London.*

1. Khera, A. V. *et al. Nature Genet.* **50**, 1219–1224 (2018).
2. Nikpay, M. *et al. Nature Genet.* **47**, 1121–1130 (2015).
3. Tikkanen, E., Havulinna, A. S., Palotie, A., Salomaa, V. & Ripatti, S. *Arterioscler. Throm. Vasc. Biol.* **33**, 2261–2266 (2013).
4. Maas, P. *et al. JAMA Oncol.* **2**, 1295–1302 (2016).
5. Lee, J. J. *et al. Nature Genet.* **50**, 1112–1121 (2018).

COMMENT

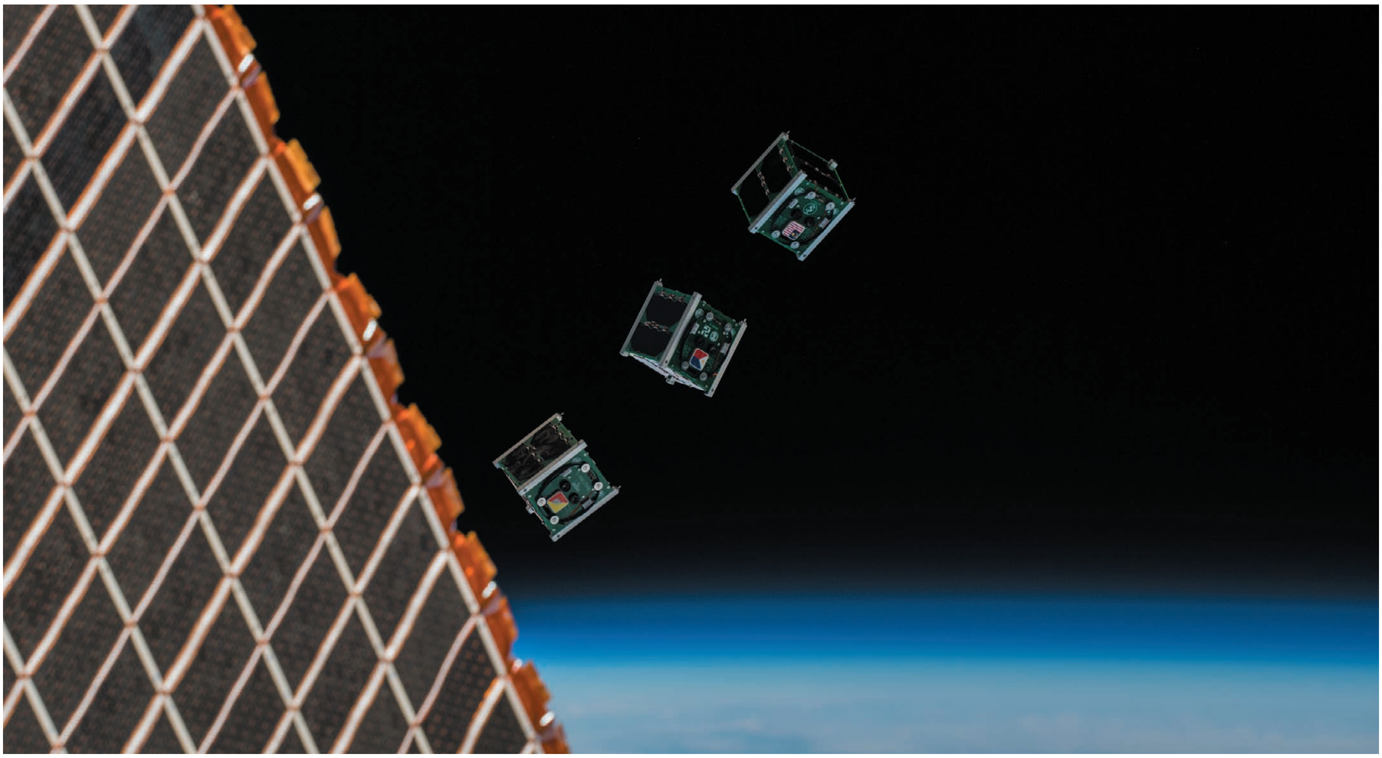
CYBERSECURITY The evidence that 'weaponized rumours' hacked 2016 US election **p.188**

SCI-FI What did Isaac Asimov and L. Ron Hubbard have in common? **p.189**

NATURAL HISTORY The sisters who made exquisite etchings of snails **p.190**

FUNDING Rewarding swagger squeezes out humility **p.192**

ESA/NASA-A. GERST



Three miniature satellites — CubeSats — launching into orbit from the International Space Station in August 2018.

Explore space using swarms of tiny satellites

Sand-grain-sized computers, self-healing materials and constellations of craft would reboot our reach, explain **Igor Levchenko**, **Michael Keidar** and colleagues.

"The first trillionaire will be made in space," US Republican Senator Ted Cruz told scientists and entrepreneurs in May at a Washington DC summit on sending humans to Mars. He could be right, but only if we rethink space technology.

The cost of launching a satellite is comparable with the value of its weight in gold. It takes thousands of dollars to send one kilogram into low Earth orbit, often ten times more than that. Returning material is even more expensive: it cost the equivalent of US$250 billion per kilogram of sample for Japan's Hayabusa spacecraft to bring back less than 1 gram of asteroid grains in 2010. The price tag for the whole mission was $250 million.

Still, space is big business. Globally, companies invested about $262 billion in 2016, mostly on using satellites for telecommunications, navigation and remote sensing[1] (see 'Lift-off'). Governments, too, spend billions — about $84 billion worldwide in 2016. More than half that ($48 billion) was from the United States, mainly for military, meteorological and communications purposes.

No one is getting much bang for those bucks. Space hardware has not kept pace with technology development and needs to be modernized. Satellites are still too bulky and expensive. Most perform only a limited set of predefined tasks. And, despite the skill and materials that went into them, they fail within decades — much more quickly than a Swiss watch.

At this rate, humans will never venture far from Earth, let alone colonize the Moon and Mars or capture asteroids.

Here we highlight three ways in which space technology needs to advance. Costs must be slashed; satellites should be small, ▶

nimble and able to repair themselves; and they should operate in swarms.

MINIATURIZATION

Satellites are shrinking. More than 800 CubeSats are now in orbit. Made from palm-sized modules, these measure about 10 centimetres across and weigh only a kilogram or so. And researchers could soon be able to package the entire 'brain' of a satellite into 1 cubic millimetre. For example, in March, IBM demonstrated a computer the size of a grain of salt, containing 1 million transistors. The smaller such devices become, the less energy they need to run, and the lighter and cheaper they are to launch.

Satellites come in two types. Passive ones need only orientation and stability control. Active ones can be manoeuvred using thrusters. Passive satellites are easiest to miniaturize. We anticipate that they could weigh in at less than 100 grams if the hardware used for controlling stability could be made less bulky. Together, thousands of these 'femtosatellites' could operate as a network.

Active satellites would take longer to shrink. As Russian poet Vladimir Mayakovsky said (of mining radium), "For every gram you work a year." They would need minute propulsion systems. Electrical techniques are most efficient. These include: microcathode arc thrusters that use electrical arcs to convert solids into plasma; electrospray systems that generate microdroplets or ions; thrusters based on field emissions that produce energetic ions; and gas-fed systems, such as miniaturized Hall-effect thrusters, in which the propellant is accelerated by an electric field.

Standard designs of tiny satellites will be needed to speed up development, production and deployment, and to save money. But the designs must be customizable so that they can, for example, support bespoke scientific instruments and protect sensitive components from heating or irradiation when necessary. Many design templates will need to be pursued at once.

Tiny satellites need small rockets to launch them. Although industry interest remains strong for large carriers such as the Falcon 9 rocket (which is capable of carrying hundreds of small satellites as well as big ones), 'microrockets' are being developed by emerging companies such as Vector Launch in Tucson, Arizona (of which one of us, J.C., is chief executive), Firefly Aerospace in Cedar Park, Texas, and Gilmour Space Technologies in Queensland, Australia. Microrockets are relatively cheap and quick to make. They weigh a few tonnes — much less than the 500-tonne Falcon 9 or 733-tonne Delta IV Heavy. Small rockets fitted with small, simple engines (that use solid propellants) could deliver dozens of CubeSats at once to low Earth orbit, potentially daily.

LONGEVITY

Before we blast thousands of small satellites or interplanetary probes into space, we must ensure that they will keep working. A swarm of unreliable satellites faulting like bulbs in a string of lights would hardly be efficient. Longevity is crucial for colonizing the Moon and Mars, where equipment failure might mean life or death.

Today's satellites are typically designed to last for between 1 and 15 years. Some space technology survives for longer: the 41-year-old Voyager 1 probe left our Solar System in 2012, but it is unlikely to send us back a message 40,000 years from now, when it is due to pass near the star Gliese 445 in the constellation Ursa Minor. Satellites disintegrate quickly because space is hostile — extremely cold, almost a vacuum and peppered with high-energy particles and ionizing radiation.

Building in redundancy can only go so far. For example, the Curiosity rover on Mars was intended to work for about 500 Martian solar days (sols). It celebrated sol 2,000 in March — although it has small breaks on at least one of its six wheels. Adding spare wheels is an obsolete approach.

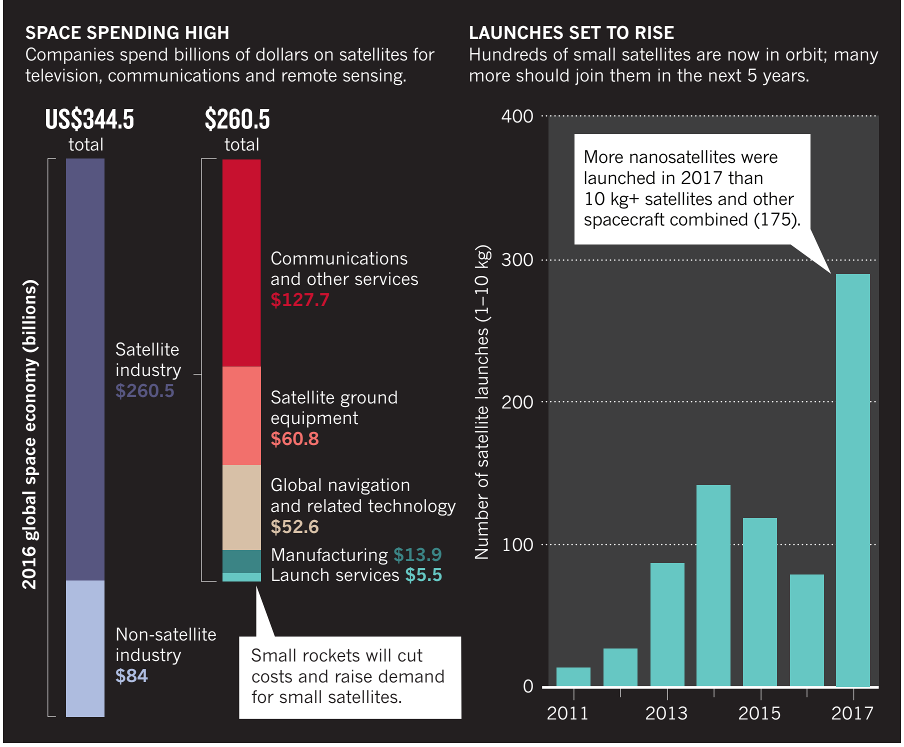
If satellites are to remain functional for a century or more, they need to be able to regenerate — as living organisms do. For example, the jellyfish *Turritopsis dohrnii* can rejuvenate almost indefinitely. Whenever it feels threatened or is injured, it reverts from its mature medusa state to the polyp state, thus beginning its life again. It can do this several times a year, depending on the environment. Some more-complex animals, such as axolotls (*Ambystoma mexicanum*), can grow new limbs, and microscopic tardigrades can survive in outer space.

Likewise, in space, human habitats, as well as tanks containing fuel and air, must be able to plug punctures and cracks autonomously. Batteries, electric generators and sensors should repair themselves when they are damaged. Some materials capable of self-healing have been developed in the lab, including flexible laminates, polyurethane composites, metallic materials and semiconducting polymers[2–4]. NASA recognized this need in its 2017 technology investment plan[5]. But a lack of collaboration between materials scientists and space technologists is slowing development.

Other types of advanced materials that are ripe for exploitation in space include durable and self-repairing lightweight and flexible structures for exploration and colonization missions. Materials with special heat properties are needed for spacecraft re-entering the atmosphere of Earth or other planets. Carbon-nanotube scaffolds, mimicking the nanoscale structures of sea shells, might increase the toughness of materials and improve ceramics. Strategies are also needed to stop cracks propagating and to prevent fatigue damage from accumulating. Environmentally friendly

LIFT-OFF

Satellites make up three-quarters of the global space economy. They are being launched in record numbers as the costs of building and putting them into orbit come down.



SOURCES: REF. 1 (LEFT PANEL); SPACEWORKS (RIGHT PANEL)

A rocket built by US firm Garvey Spacecraft (now part of Vector Launch) carried four CubeSats in 2013.

NASA/DIMITRI GERONDIDAKIS

materials are desirable.

Researchers need to explore adaptation. Spacecraft might have to deal with the unexpected, such as grabbing irregularly shaped asteroids or handling other satellites for repair missions. Adaptable grippers, made from elastic or intelligent materials, need to be designed. Eventually, we'll need fully self-repairing space platforms, including propulsion systems, power plants, life-support systems and scientific instruments. Building even a prototype demands major breakthroughs and new ways of working.

NETWORKING

Instead of building one satellite to perform a single task, constellations of thousands of satellites have much broader potential. Their instruments can operate together as if they were on a much larger platform. For example, the five satellites that now make up the Afternoon Train constellation monitor clouds, aerosols and greenhouse and other gases in Earth's atmosphere to provide 3D reconstructions of climate and weather patterns and atmospheric pollution. In the CANYVAL-X mission, two CubeSats fly in formation to develop techniques that will help to study the Sun (one is equipped with a microcathode arc thruster).

Many configurations are possible — from trains of satellites following one another along the same orbit, to evenly spread networks watching Earth's entire surface (and, in future, maybe also those of the Moon and Mars). The constellation's shape can be adjusted. Several networks can be linked together virtually, to increase their power, resilience and responsiveness. Some satellites might be tooled to repair and adjust others.

Swarms of miniature satellites should be cheap and quick to deploy. Thousands could be released from a large central satellite in orbit. Swarms able to receive and send signals and perform basic logic operations could be combined with clusters of fewer, larger, more-complicated and manoeuvrable satellites that act as communications or analysis hubs.

Ultimately, constellations might behave like a neural network or artificial intelligence. Collective properties could be exploited, such as self-organization, transformability, self-learning and simultaneous sensing over a large area — as in the clouds of microscopic, interacting robots envisaged by Polish science-fiction writer Stanisław Lem in his 1964 book *The Invincible*.

So far, only tens of satellites have been strung together. The Global Positioning System (GPS) satellite constellation, for example, requires about 30 operational satellites for reliable global coverage. Efforts are afoot to increase the numbers. In Japan, Hokkaido and Tohoku universities have partnered with other organizations to send 50 microsatellites into space by 2050 (each weighing about 50 kg) to trace the aftermaths of natural disasters. The Iridium telecommunications network is being boosted to contain around 80 satellites.

By the mid-2020s, the company SpaceX intends to launch 12,000 small satellites to set up Starlink, a space-based Internet network. Two prototype Starlink satellites were launched in February, and the network may begin operating as soon as 2020. The communications company OneWeb aims to ensure affordable global access to Internet services through a constellation of 600–2,000 small satellites (up to 200 kg), with the first slated to be launched as early as December. Boeing's proposed constellation of 1,300–3,000 communications satellites is another example.

However, the satellites in most of these constellations are controlled from the ground. To operate efficiently, constellation units need to be able to communicate with each other and to adjust their positions and orientations in real time.

NEXT STEPS

Experts in advanced nano- and metamaterials and propulsion need to collaborate more to develop self-healing, regenerative materials for space applications. These range from composite materials for human habitats and large inflatable structures, to ultra-hard ceramics for thrusters. Micro-thrusters need to be more efficient and reliable. Unconventional systems, such as thin-film and 3D-printable thrusters, also need attention. This will require a continuing dialogue between materials scientists, propulsion experts and robotics specialists, which should begin in conferences on material advances in space technology, such as the International Conference on Micropropulsion and CubeSats (www.micropropulsion.org). Commercial companies will reap the benefits, and should contribute to the millions of dollars the research teams will need.

Mass-production methods must be optimized for delivering constellations of thousands of satellites. Additive manufacturing techniques such as 3D printing are lowering the costs of custom satellites. Production methods must be factored in when designing space technologies. Designs of auxiliary systems such as launch pads, thruster platforms and power and control systems must be standardized.

In addition, policymakers and lawyers need to develop an international legal framework for operating large constellations. For example, licences and permissions are needed to launch craft. Communication frequencies and orbits need to be assigned. The de-commissioning and removal of satellites at the end of their working lives must be coordinated internationally. Insurance needs to be established for losses from delays in the deployment of satellites, as happened for the Iridium NEXT mission to upgrade its constellation.

It is too soon to say whether the space economy will become profitable. But central to that economy will be the coming constellations of tiny satellites. ■

Igor Levchenko *is a research scientist at the National Institute of Education, Nanyang Technological University, Singapore.* **Michael Keidar** *is professor of engineering and founding director of the Micropropulsion and Nanotechnology Laboratory at George Washington University, Washington DC, USA.* **Jim Cantrell, Yue-Liang Wu, Hitoshi Kuninaka, Kateryna Bazaka, Shuyan Xu.**
e-mail: levchenko.igor@nie.edu.sg

1. Bryce Space and Technology. *Global Space Industry Dynamics: Research Paper for Australian Government, Department of Industry, Innovation and Science* (Bryce Space and Technology, 2017).
2. Bauer, S. & Kaltenbrunner, M. *Nature* **539**, 365–367 (2016).
3. Kim, J. T. *et al. Sci. Rep.* **8**, 2120 (2018).
4. Oh, J. Y. *et al. Nature* **539**, 411–415 (2016).
5. NASA. *NASA Strategic Technology Investment Plan 2017* (NASA, 2017).

Social-media posts spread false information about Democratic candidate Hilary Clinton during the 2016 US presidential election campaign.

CYBERSECURITY

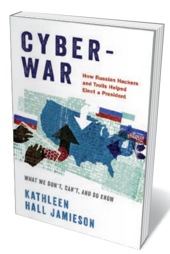
Hacking the presidency

Alexander Klimburg lauds a study probing Russia's impact on the 2016 US elections.

Late in 2016, then-US President Barack Obama mused in an interview with *The New Yorker* magazine that he had probably been elected because his campaign had begun before the old media order collapsed. Communication scientist Kathleen Hall Jamieson's illuminating, timely *Cyberwar* is a major step forward in trying to understand the 'new' media order — and how open this digital landscape is to malicious exploitation. Jamieson's focus is Russian involvement in the 2016 presidential elections; her implicit conclusion is that, very probably, it had a major role in Donald Trump's surprise win.

Jamieson provides perhaps the first authoritative collection and synthesis of the copious amounts of open data surrounding the 2016 attack. She draws on several published US intelligence accounts, indictments, media reports and a wealth of research in communication studies to reveal Russia's part in Trump's victory, and how much it depended on the digital propagation of fantasy narratives to mobilize or demobilize supporters. Russian "discourse saboteurs" (trolls) in St Petersburg "farms" and beyond were able to do two things: exploit weaknesses in social-media platforms, and count on coldly cynical US fellow travellers willing to disseminate false rumours. Unwittingly, the mainstream media, Jamieson reveals, played a key part.

For Russia, Jamieson shows, it was win–win–win. If successful, it would get a candidate it thought useful. If not, it would have seeded the idea that Hillary Clinton had rigged the election, making it difficult for her to govern. Either way, through unrestrained intervention, Russia would advance its overriding narrative — that information and speech are weapons that need to be controlled.



Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know
KATHLEEN HALL JAMIESON
Oxford University Press (2018)

In addition to her academic position at the University of Pennsylvania in Philadelphia, Jamieson is co-founder of FactCheck.org, a non-partisan website and project of the university's Annenberg Public Policy Center. It describes itself as a "'consumer advocate' for voters that aims to reduce the level of deception and confusion in U.S. politics". *Cyberwar* is, appropriately, highly circumspect on what is known and not known about Russian interference in this election.

The result is perhaps the clearest-cut glimpse of what an information war looks like. Ultimately, it helps to explain how 80,000 Facebook posts, 131,000 tweets and 1,000 YouTube videos created by at least one group of Russian operatives might have thrown the election. These activities, and sharing by other users, reinforced each other. In the end, the trolls reached more than 126 million Americans through Facebook alone.

Jamieson divides her analysis into four key parts, reflecting the intentions of Russian messages: priming, framing, agenda setting and contagion. In nearly all cases, the trolls were able only to amplify or build on the system's existing weakness.

The priming and framing of many election themes preceded the campaign, and were reinforced by troll activity. Clinton's characterization as a "dishonest" woman was an established Republican framing, as Jamieson shows. During the campaign, it was constantly encouraged by casting Clinton as a dissembler — for instance, over her use of a private e-mail server, and in relation to false rumours pumped out continuously by trolls on social media, under assumed names. These stories and memes were

ALEX WONG/GETTY

shared through trusted social connections on Facebook, YouTube, Twitter and other platforms, exploiting the "two-step flow" of propagation, in which interpersonal relationships increase the traction of a message.

That flow also helped to set an agenda in the mainstream media. Jamieson shows how much news reportage was triggered by uncritical tracking of Twitter and Facebook memes. Finally, the contagion effect did much to ensure that even attempts to dismiss the more ludicrous conspiracy theories meant that negative associations still clung to Clinton. Facebook became a "contagion machine", Jamieson writes. Its algorithms quickly learnt that the best way to retain users was to keep them angry and afraid — responses that troll messages were designed to elicit.

As Jamieson writes, the trolls aimed strategically to direct attention to hot-button issues such as illegal immigration or police brutality. Exploiting the two-step flow, the trolls gained traction in niche groups by pretending to be extremists in both left-wing and right-wing camps and sending out messages ranging from exaggerations to complete fictions. As Senator Mark Warner (Democrat, Virginia) of the Senate Intelligence Committee recounts, these efforts were largely directed at demobilizing possible Clinton voters. Meanwhile, trolls tried repeatedly to incite violence, attempting to organize at least 129 rallies on both left and right — some at the same time and place, with the clear intent that they should clash.

The media's frames of choice prevented the full implications from sinking in — for instance, casting e-mails hacked from the Democratic National Committee and published in 2016 as 'leaked correspondence'. And the media inadvertently aided counter-messaging that protected Trump from bad press (such as recordings of him speaking lewdly while filming for the *Access Hollywood* programme) through the timing of reports, even distracting from US government announcements that a Russian disinformation campaign was under way.

"Trolls aimed strategically to direct attention to hot-button issues such as illegal immigration."

Cyberwar is all the more powerful for what it is not. It is not a book of international politics or warfare. Its title is likely to displease those who think it might inadvertently support those actors (such as Russia) who wish to cast information warfare as 'war'. It does not attempt to portray the full landscape of this new, cyber-enabled cold war. It describes only part of the new conflict paradigm, which also includes Russia's preparations for 'real', critical-infrastructure-crashing cyberwar, along with the slow and steady erosion of the Western alliance, democracies and international law writ large — all in an attempt to fulfil a zero-sum world view in which Russian greatness can be (re)achieved only by vanquishing the country's implacable foes.

Indeed, Jamieson plays little heed to accusations that the actual electoral system — voting machines and voter registries — might have been tampered with. She concludes (rightfully, in my view) that if they had been, the manipulation would probably represent only a fraction of the votes 'stolen' through troll activity. In the end, Jamieson's final analysis is clear, if not explicit: Russian trolls must have swung many more votes than the 78,000 in 3 crucial states that constituted Trump's winning Electoral College margin. Indeed, the reader is left with the distinct impression that the number of affected votes was probably orders of magnitude higher.

Cyberwar provides a convincing model of how the old Soviet 'active measures' of propaganda, honed throughout the twentieth century, can be enacted with great effect under the new media order. Most importantly, Jamieson specifies the roles of complicit citizens and an unwitting media. By showing that modern Western democracy has a significant existential challenge, she has set us on the path to help patch it — if only we are able to move fast enough. ■

Alexander Klimburg *is a senior non-resident fellow at the Atlantic Council in Washington DC and an affiliate of the Berkman Klein Center at Harvard University in Cambridge, Massachusetts. He is the author of* The Darkening Web.
e-mail: alexklimburg@hcss.nl

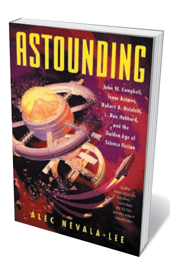
SCIENCE FICTION

How science fiction grew up

Rob Latham savours the convoluted tale of four men who reshaped the genre.

Alec Nevala-Lee's *Astounding* is a fascinating collective portrait of four men who, together and apart, helped to shape modern science fiction. They were the legendary, irascible John W. Campbell Jr, long-time editor of the magazine *Astounding Science Fiction* (later *Analog*), and three of his key writers. Isaac Asimov and Robert A. Heinlein became giants of the genre. L. Ron Hubbard, by contrast, was a prolific purveyor of pulp fiction (and future founder of the Church of Scientology).

Under Campbell's editorship, *Astounding* was transformed during the late 1930s and 1940s from a showcase for space-opera schlock into a serious venue for futuristic extrapolation, often written by professional scientists such as Asimov, a biochemist, and electronics engineer George O. Smith. That era has become known as science fiction's golden age. Nevala-Lee — himself a science-fiction writer — delivers a compelling account of its hopeful rise and ignominious fall.



Astounding: John W. Campbell, Isaac Asimov, Robert A. Heinlein, L. Ron Hubbard, and the Golden Age of Science Fiction
ALEC NEVALA-LEE
Dey Street (2018)

Pivotal in this trajectory was the massive, lingering impact of the Second World War on the magazine and its stable of authors, several of whom were drawn into military research. Asimov, Heinlein and fellow *Astounding* regular L. Sprague de Camp tested war materials at the Philadelphia Navy Yards in Pennsylvania from 1942. Campbell, under the aegis of the University of California's Division of War Research, led a team of authors revising technical manuals for military use. He also joined Heinlein and de Camp in brainstorming unconventional responses to kamikaze attacks, such as detecting approaching aeroplanes using sound.

Despite knowing that publishing stories treating potential new forms of military technology would run afoul of the wartime censors, the ever-obstinate Campbell did just that in March 1944. Cleve Cartmill's 'Deadline' depicted the invention of a nuclear bomb using isotopes of uranium. Campbell, a trained physicist who strongly suspected the government was working on such a weapon, fed technical details to Cartmill, who set the tale on another planet. (Cartmill slyly called the warring aliens Sixa and Seilla, Axis and Allies spelt backwards.)

Unsurprisingly, the story drew the ▶

Astounding Science Fiction's cover for May 1947.

GRANGER HISTORICAL PICTURE ARCHIVE/ALAMY

attention of the national Counter-intelligence Corps, which suspected a leak from the Manhattan Project; swathes of the personnel at the project's site in Los Alamos, New Mexico, were science-fiction fans. Campbell was aggressively interviewed by an intelligence agent, Cartmill's personal correspondence was put under surveillance, and *Astounding* came close to having its mailing privileges revoked. After the war, Campbell often cited the incident to demonstrate the genre's prophetic nature — its capacity to project a convincing fictional future from known scientific facts.

Indeed, the unprecedented technological advances of the war fuelled the public taste for science and technology, in turn raising the cultural status of science fiction. The late 1940s and 1950s were a boom time for the genre. That boosted the stock of *Astounding*, which came to specialize in stories of nuclear conflict and crisis. It also led to the rise of competing titles such as *Galaxy* and *The Magazine of Fantasy & Science Fiction*, as well as an expansion of the science-fiction book market. Campbell's talent began to be poached.

Nevala-Lee carefully traces the rifts that developed in the core group, largely prompted by Campbell's increasing fondness for pseudo-scientific ideas such as the Dean drive (proposed by inventor Norman Dean, who claimed it could produce thrust without a reaction — in violation of the laws of motion).

More generally, Campbell had always been obsessed by the possibility of a truly scientific psychology, which he believed would have predictive power along the lines of the fictional science of psychohistory in Asimov's Foundation series. So when Hubbard, in the late 1940s, shared ideas that later became his 'self-help system' Dianetics, Campbell took the bait. Hubbard's vision of superpowers purportedly lurking in everyone — once they had gone through an 'auditing' process and emerged as 'clears' — gripped Campbell, and he helped Hubbard to market his 1950 book *Dianetics*. Nevala-Lee argues that a lingering messianism at the heart of science fiction — its "persistent dream of an exclusive society of geniuses" — helped to propel Hubbard's movement, which became Scientology. Numerous sci-fi authors embraced Dianetics, submitting to auditing or even becoming trained auditors; A. E. van Vogt briefly abandoned his writing career to run a chapter in Los Angeles, California.

Hubbard's gift for the hard sell was pivotal, and Nevala-Lee's portrait of him as a paranoid narcissist and skilled manipulator is scathing. However, Campbell is also sharply scrutinized for his role in midwifing and unleashing *Dianetics*. Heinlein and Asimov were repelled by what they saw as an uncritical embrace of quackery, and took refuge in newer, often more lucrative markets. The book's final chapters detail the steady decline of the magazine into a second-rank publication, and Campbell (who died in 1971) into a reactionary crackpot with racist views.

Although much of the story outlined in *Astounding* has been told before, in genre histories and biographies of and memoirs by the principals, Nevala-Lee does an excellent job of drawing the strands together, and braiding them with extensive archival research, such as the correspondence of Campbell and Heinlein. The result is multifaceted and superbly detailed. The author can be derailed by trivia — witness a grisly account of Heinlein's haemorrhoids — and by his fascination for clandestine love affairs and fractured marriages. He also gives rather short shrift to van Vogt, one of Campbell's most prominent discoveries and a fan favourite during *Astounding*'s acme, whose work has never since received the attention it deserves.

These quibbles aside, the book is a rich, gripping cultural and historical study of how a small cadre of talents in a minor commercial genre became some of the most influential figures of the second half of the twentieth century. ■

Rob Latham *is the editor of* The Oxford Handbook of Science Fiction *and* Science Fiction Criticism. *For 20 years, he was a senior editor of the journal* Science Fiction Studies.
e-mail: rob@lareviewofbooks.org

NATURAL HISTORY

Scientific artistry of the Lister sisters

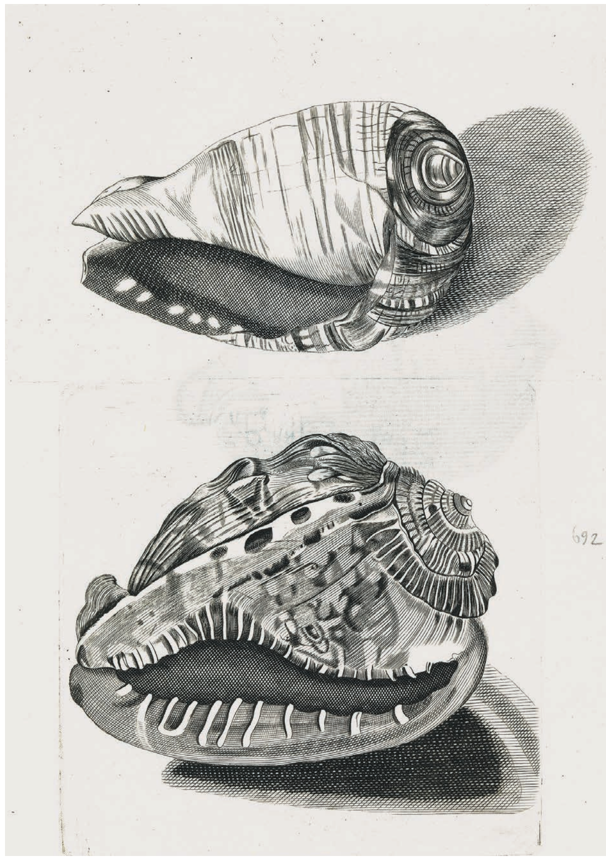
Beth Fowkes Tobin applauds a book on a gifted family of early-modern naturalists.

Between 1685 and 1692, Martin Lister — a noted British physician and naturalist — published *Historiae Conchyliorum*, a significant study of molluscs filled with hundreds of beautiful illustrations of all known shells. The illustrators were Lister's daughters Anna and Susanna. How these drawings and etchings came into being in an era that excluded women from formal scholarship is meticulously shown in *Martin Lister and his Remarkable Daughters*.

Historian Anna Marie Roos marshals her considerable talents as a researcher to recover the story of how Lister's daughters learnt to draw and etch scientifically accurate natural-history illustrations. Records of women's scientific work from this time are scant; naturalist and illustrator Maria Sibylla Merian's spectacular drawings of Surinam's insects are among the rare surviving examples.

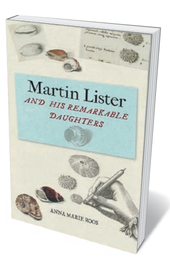
As Roos relates, Susanna and Anna Lister were in their teens when their father enlisted their services as illustrators for his ambitious project. They spent nearly a decade working on it, an amazing feat noted by Lister's friend Edward Lhwyd, naturalist and keeper of the Ashmolean Museum in Oxford, UK. In teaching his daughters how to draw, Lister also taught them how to see animal and plant specimens as a scientist would. He may have sat with his daughters while they drew the shells, to point out characteristics key to classification. (As he noted in another context, such supervision was important to ensure that "the excellent artist did not merely … express his own personal conception".)

Lister also instructed Susanna and Anna in etching and engraving, skills rarely taught to women at the time, because they were viewed as arduous and dangerous. Engraving demanded physical strength to cut the surface of the copperplate; etching, the use of hazardous nitric acid to dissolve away the metal.



Two engravings of shells from *Historiae Conchyliorum*.

The Lister sisters may have been the first women to use microscopes in producing images — for *Historiae Conchyliorum* and to accompany letters published in the Royal Society's journal, *Philosophical Transactions*. Two of these, on wood grain and salt crystals, were authored by the pioneering Dutch scientist and microscopist Antonie van Leeuwenhoek. Anna may also have learnt to dissect specimens. Annotations in her notebook indicate as much, and among her original drawings for the Listers' opus is a "depiction of a brachiopod gill and dissected mollusc penises". Anna also drew illustrations of the bodies of living snails for the final volume — an innovation that would not be repeated until the mid-eighteenth century, when French conchologists turned their attention to the mollusc itself, instead of its shell.



Martin Lister and his Remarkable Daughters: The Art of Science in the Seventeenth Century
ANNA MARIE ROOS
Bodleian Library (2018)

Roos portrays these extraordinarily talented young women as beneficiaries of their polymath father, whose indefatigable curiosity about the natural world drove his achievements, and who gave his daughters unusual latitude in pursuing the art and science of conchology. Nearly half the book is a biography of Lister as a physician to London's elite, popular travel writer and vice-president of the Royal Society. He was the first serious arachnologist and conchologist in Britain, and an expert on viticulture. He also dabbled in chemistry, pharmacology and mathematics, and contributed to theories about the age of Earth through his study of fossilized shells.

Roos places Lister at the centre of the movement towards observational empiricism in studying the natural world, as the nascent discipline shifted in focus from exotica to a more systematic gathering of data, locally and globally. Active in many scientific networks, he had contacts ranging from local York engravers, chemists and printers to luminaries such as naturalist James Petiver, physician and collector Hans Sloane and naturalist John Ray, all of whom delivered specimens, drawings, books and advice (see H. Nicholls *Nature* **545**, 410–411; 2017). Sloane, for instance, lent the Listers Jamaican shells for their work.

Roos also explores the Listers' later legacy, in particular the republication of *Historiae Conchyliorum* in 1770, allowing broader access to an authoritative work. Emanuel Mendes da Costa, author of *Elements of Conchology* (1776) and *British Conchology* (1778), complained to a friend that the Listers' volumes were "very scarce" and not to be found in booksellers' shops; he was particularly eager to purchase a reissue.

Roos describes the archival afterlife of Anna and Susanna's works and equipment, complete with tales of treasures lost and found. The copperplates, thought to be lost, were rumoured to be housed in tea chests at Oxford, until Roos tracked them down in the Bodleian Library.

The colour plates in Roos's book are gorgeous, especially those from the notebooks and sketchbooks, their beauty heightened by their ephemerality. Martin Lister's drawing of a strawberry finch reveals a skilled illustrator who "performed a type of embodied empiricism". Even more stunning, some of the reproduced illustrations are juxtaposed with photographs of the very shells they depict. Roos thus showcases a material legacy central to the history of how early-modern scientific books were produced.

Martin Lister and his Remarkable Daughters is lucid and on occasion surprisingly funny. Most of the time, Roos keeps her narrative threads meshed, interweaving the separate achievements of father and daughters with the trajectory of their opus. But I wished for less biographical detail on Martin Lister's youth and education. (This material is readily available in Roos's 2011 monograph *Web of Nature*.) Those eager to learn about Anna and Susanna and their scientific and artistic legacy will be delighted by the final two chapters and the photographs of their artwork.

As for their personal lives, details are scarce; but Roos has scrounged a few tantalizing titbits. After their epic stint on the book ended, Susanna married one Gilbert Knowler, becoming his third wife. Less is known about Anna, but Roos has uncovered the possibility that she married John Bristow in 1701 against her father's wishes, which would explain why she was not mentioned in his will. Yet, however rich the biographical detail on Martin Lister, the sisters' exquisite scientific contribution tells a story of its own.

Roos is to be congratulated on recovering an important episode in the intertwined history of art and science in the early-modern period, the history of scientific-book production and the hidden role of women in the history of science. ■

Beth Fowkes Tobin *is professor of English and women's studies at the University of Georgia in Athens, and has published widely on eighteenth-century British natural history, including the books* The Duchess's Shells *and* Colonizing Nature.
e-mail: btobin@uga.edu

Correspondence

Funding: practices risk promoting bias

Funding processes seem to us to be rewarding only particular types of scientist. This is leading to discriminatory practices in the very institutions that encourage scientists to overcome their implicit biases when making decisions and assessments.

Drawing examples from biomedicine, UK funding initiatives are increasingly calling for applications from investigators who feel they are potentially future leaders who can make a leap, tackle a grand challenge, be transformative and advance a unique, game-changing strategic vision. Such wording risks discouraging more-modest scientists and those patiently pursuing slowly unfolding advances.

Interviews that are designed to seek out such 'winning' qualities could select against those scientists who might be unnerved by a daunting committee. By extension, academic institutions must recruit scientists who fit these norms if they are to succeed in today's competitive funding climate.

Efforts to promote diversity in science will fail if the exemplar of a successful scientist is so narrowly defined. We need more-inclusive hallmarks of performance, as well as equality legislation and training.

Wendy Bickmore, **Sarah Cunningham-Burley**, **Margaret Frame** *University of Edinburgh, UK.*
wendy.bickmore@igmm.ed.ac.uk

Funding: gamble on radical proposals

The competition to secure funding can deter applicants from submitting radical research proposals, despite their potential for dramatic advance. At University College London (UCL), we have been running a programme for ten years that bypasses conventional funding mechanisms, using our own resources to open up new and unpredictable lines of enquiry.

A grant-application system such as that used today would probably have denied support to many of the twentieth-century scientists who fundamentally changed the ways we think. For example, molecular biologist Oswald Avery and his colleagues disproved the widely held belief that the genetic molecule was a protein (O. T. Avery *et al. J. Exp. Med.* **79**, 137–158; 1944).

UCL took its lead from British Petroleum's Venture Research Unit (1980–93), which awarded funding to a handful of applicants with radical ideas — simply on the basis of face-to-face discussion.

Despite vetoes by peer reviewers, the unit supported academics such as Ken Seddon, who became the United Kingdom's most cited chemist for his work on ionic liquids, and Steve Davies, who set up a company to further his research into molecular architecture and chiral selection. The company sold in 2000 for £316 million (then about US$200 million) — some 15 times the unit's total outlay on venture research.

Universities should follow UCL's lead and use their own resources to set up similar initiatives.

Don Braben *University College London, UK.*
don.braben@btinternet.com

Keep groundwater clear of pesticides

Pesticide residues from Denmark's intensive-farming industry are contaminating the country's groundwater, which is used exclusively as its source of tap water (see, for example, go.nature.com/2iumpdx). This is despite the governments's raft of protection measures that have been in place since 1994 (see go.nature.com/2xhinf7).

Pesticide residues in drinking water are a threat to public health. They can compromise neuroendocrine development in unborn and newborn children and can lead to chronic kidney diseases in later life (X. Xu *et al. Nature Rev. Nephrol.* **14**, 313–324; 2018), as well as to other, unforeseeable effects.

Pesticides therefore need to be removed at the waterworks before consumption — a process that is economically and environmentally costly. And it is uncertain whether current technology can remove all such residues (P. J. J. Alvarez *et al. Nature Nanotech.* **13**, 634–641; 2018).

We call for greater political accountability and better management of the country's groundwater. In our view, areas where groundwater is abstracted should be protected against pesticide use and farmers should receive economic compensation.

Without such measures, Denmark could end up losing its role in setting the agenda for sustainable use of pesticides through European Union directives, the United Nations Environment Programme and the Stockholm Convention on Persistent Organic Pollutants.

Christian Sonne, **Martin Hansen** *Aarhus University, Roskilde, Denmark.*
Aage K. Olsen Alstrup *Aarhus University, Aarhus, Denmark.*
cs@bios.au.dk

Mouse avatars guide immunotherapy

We think your discussion on the use of mice with human tumours as cancer models is too pessimistic (*Nature* **560**, 156–157; 2018). These mouse 'avatars' can now be armed with human immune cells and are already providing promising insights into immunotherapies (Y. Choi *et al. Exp. Mol. Med.* **50**, 99; 2018).

One example is a personalized mouse model we developed for melanoma. Here, the tumour and immune cells come from the same individual and the response of the mouse to immunotherapy matches that of the patient (see H. Jespersen *et al. Nature Commun.* **8**, 707; 2017).

Difficulties in getting some human grafts to grow successfully in mice could hinder the widespread application of avatar techniques in routine cancer care. Melanoma xenografts are unusual in that they engraft and grow fast enough to support the initiation of immunotherapy in patients. For ethics reasons, however, avatars are better suited to clinical research, for example, to screen patients' suitability for trials.

Jonas A. Nilsson, **Roger Olofsson Bagge**, **Lars Ny** *University of Gothenburg, Sweden.*
jonas.a.nilsson@surgery.gu.se

Antibiotic resistance pre-dates penicillin

Clinical antimicrobial resistance was first reported four years before Alexander Fleming's discovery of penicillin in 1928. The antimicrobial in question was known as Salvarsan (S. Silberstein *Arch. Derm. Syph.* **147**, 116–130; 1924).

An antibiotic was originally defined as an agent that microorganisms produce to kill competing bacteria (S. A. Waksman *Mycologia* **39**, 565–569; 1947). This has been extended to include synthetic drugs, including sulfonamides and quinolones. Salvarsan was one such drug, from a group of compounds known as arsphenamines. It was used to treat syphilis from 1910 until the 1940s, when penicillin took over because it was more readily available, safer and more effective.

Bacterial resistance to Salvarsan started to emerge about halfway through that period, despite the drug's limited use by comparison with modern antibiotics. The 1924 paper was cited by several groups during the 1930s (see, for example, W. Beckh and G. V. Kulchar *Arch. Derm. Syphilol.* **40**, 1–12; 1939), but has long since been forgotten.

Dov Stekel *University of Nottingham, UK.*
dov.stekel@nottingham.ac.uk

HUMAN GENOMICS

Biobank for the masses

UK Biobank contains a wealth of data on genetics, health and more from 500,000 participants. A detailed overview of the biobank and an analysis of its brain-imaging data show the value of this resource. SEE ARTICLES P.203 & P.210

NANCY COX

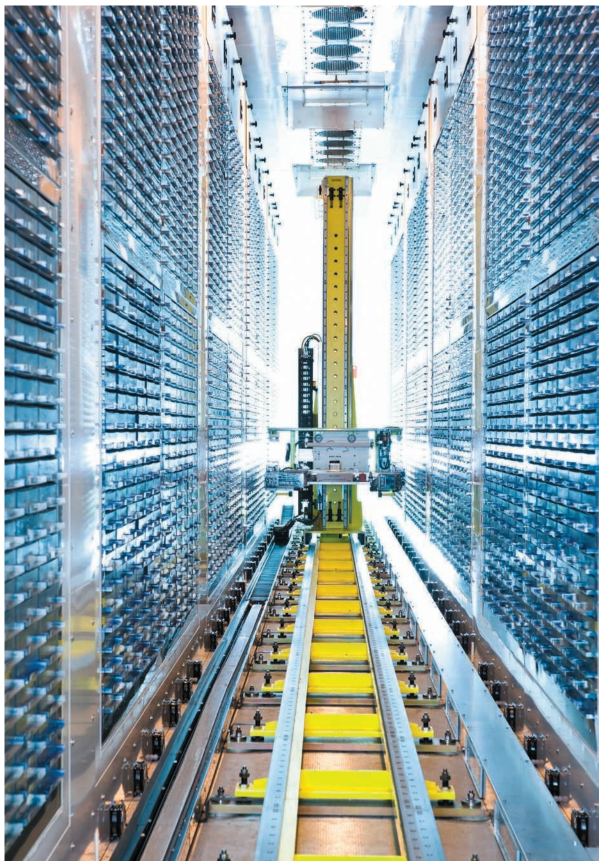
Huge sample sizes are often needed to discover the genetic variants that contribute to disease. Meta-analyses of many genome-wide association studies (GWAS), which test for such links, are now beginning to search for associations between DNA variants and common diseases in more than one million individuals[1]. But perhaps equally important is detailed clinical and biological information about the participants, which enables researchers to better test for more associations — including those that give insight into disease mechanisms. Writing in *Nature*, Bycroft *et al.*[2] and Elliott *et al.*[3] describe a huge resource called UK Biobank that marries large-scale genomic and detailed clinical data for 500,000 people. The biobank promises to aid the discovery of relationships between genome variation and common human diseases, and to improve our understanding of the mechanisms that underlie those associations.

As Bycroft *et al.* describe, UK Biobank's 500,000 participants donated urine, saliva and blood samples (Fig. 1), which were used for genetic analysis and evaluated for known biomarkers of disease. The participants were aged between 40 and 69 when they were recruited to the study between 2006 and 2010. This age range meant that participants would be at risk of developing common diseases of adulthood. The volunteers filled out thorough questionnaires about a wide range of factors, including family disease history, demographic background and lifestyle. They also gave consent for researchers to access electronic health-record data. Subsets of participants underwent more-comprehensive examinations, including extensive imaging and lung-function studies.

The sample size of this resource, combined with the breadth of data that have been collected — and that will continue to accrue as the participants age — is unprecedented, as is the generosity of the project's data-sharing plan. From the beginning, the intent has been to share the data in their entirety with any health researcher. As a consequence, thousands of scientists from all over the world have been doing research on these data since July 2017.

In 2007, the Wellcome Trust Case Control Consortium published a landmark study[4] that set the standard for how GWAS should be performed and the resulting data shared, greatly influencing how GWAS were conducted. Similarly, Bycroft *et al.* provide a wealth of detail on how they designed their study and analysed the resulting genetic data. As such, their paper promises to influence a new generation of data scientists.

The work is a vivid reminder that data generation is perhaps the least challenging aspect of big-data science. The researchers used an array-based approach to determine nucleotide variation at more than 800,000 genomic sites, and then imputed variation at millions more sites. But the scale of the data meant that both the design of this 'genotyping' and the subsequent quality-control analysis needed to be wholly reconceived from methods used for smaller studies. Moreover, much of the software used needed to be substantially revised to achieve reasonable computing times. Software is being made available to scientists, along with the full results of the authors' preliminary GWAS and phenome-wide association studies, the latter of which analyse associations between the entire range of traits logged in the biobank and a single genetic variant.



WELLCOME IMAGES

Figure 1 | Biological samples in a storage freezer at UK Biobank. Two papers[2,3] describe the set-up of the biobank and analyse some of its data.

Bycroft *et al.* conducted several analyses to demonstrate that the data they collected would yield appropriate results in association studies. For example, they analysed a genomic region that harbours several human leukocyte antigen (HLA) genes, which have a role in distinguishing foreign cells and particles from those of our own bodies. It is well established that many variants in these genes are associated with common diseases[5]. The authors confirmed that the HLA types imputed from their genotype data have the expected associations with disease, validating both the genotype and disease data used in the study. The group also performed GWAS to identify genetic variants associated with differences in height — again, their results matched those from GWAS meta-analyses that used independent samples.

Whereas Bycroft and colleagues detail how the biobank's genome data were generated and highlight the quality of the data, Elliott *et al.*[3] give us a preview of how these data can be used to drive discovery and to probe the mechanisms underlying genetic associations with disease.

The authors focused on brain-imaging data from more than 8,400 UK Biobank participants. These data were processed to generate a list of thousands of image-derived phenotypes (IDPs) — traits related to brain structure or function that can be identified through images.

Elliott and colleagues investigated associations between IDPs and genetic variants.

The authors' analysis provides new data on the heritability of IDPs, for instance demonstrating that the volume of a given brain region is more heritable than are measurable aspects of its function. Reassuringly, these results generally replicate those from previous studies that analysed a small subset of the IDPs in a greater number of individuals[6,7].

Elliott *et al.* also demonstrated how GWAS on IDPs can be combined with the results of GWAS on neurological and psychiatric disorders as a way to gain insight into possible mechanisms of disease. For instance, they showed that variation at a particular genomic region that has previously been associated with risk of schizophrenia is also associated with certain aspects of brain volume, pointing to a possible mechanism for how and why variants in this region might be associated with disease risk. This work is just a tantalizing teaser of how much more we will learn once 100,000 UK Biobank participants have undergone brain imaging — a project that should be completed by 2020.

The excitement about the opportunities to advance human genetics using UK Biobank is palpable. Most of the variants incorporated in the biobank's database are common, but sequence data being generated to interrogate rare variants will soon be available to investigators. The size and breadth of the resource, coupled with the many related individuals who have donated their samples to this huge database, should enhance our ability to study the consequences of rare variation on a scale we could not have imagined just a few years ago.

The generosity of the United Kingdom in sharing this resource with the rest of the world is a shining example of the value of investing in the greater good. It can be challenging to make large-scale clinical data publicly available, because of privacy concerns and the difficulties inherent in removing all potentially identifying information from electronic health records. Nevertheless, scientists benefit hugely from the broad availability of all of these data sets. The US National Institutes of Health initiative All of Us is being designed to be broadly available to the scientific community. We can celebrate the United Kingdom's generosity best by emulating it. ■

Nancy Cox *is at the Vanderbilt Genetics Institute, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee 37232–0700, USA.*
e-mail: nancy.j.cox@vanderbilt.edu

1. Evangelou, E. *et al. Nature Genet.* **50**, 1412–1425 (2018).
2. Bycroft, C. *et al. Nature* **562**, 203–209 (2018).
3. Elliott, L. T. *et al. Nature* **562**, 210–216 (2018).
4. The Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
5. Moutsianas, L. & Gutierrez-Achury, J. *Methods Mol. Biol.* **1793**, 111–134 (2018).
6. Hibar, D. P. *et al. Nature* **520**, 224–229 (2015).
7. Fornage, M. *et al. Ann. Neurol.* **69**, 928–939 (2011).

DEVELOPMENTAL BIOLOGY

A dual origin for blood vessels

Contrary to previous assumptions, it seems the cells that line blood vessels are derived from more than one source. In addition to their known developmental path, they can arise from progenitors of embryonic blood cells. SEE ARTICLE P.223

M. LUISA IRUELA-ARISPE

Blood-cell lineages and the endothelial cells that line the interior of blood vessels have an intertwined biology and interrelated embryonic origins. Our current knowledge indicates that endothelial cells differentiate directly from one of the three main cell layers of the early embryo (the mesoderm), and that a subset of endothelial cells subsequently gives rise to haematopoietic stem cells (HSCs)[1,2], from which adult blood cells derive. On page 223, Plein *et al.*[3] reveal a second origin for endothelial cells, and refine our understanding of the relationship between the endothelial and blood lineages.

Transient embryonic populations of red blood and immune cells arise early in development, before the emergence of HSCs, from precursor cells called erythro-myeloid progenitors (EMPs). In line with the model that mesoderm gives rises to endothelium, which in turn gives rise to blood, EMPs originate from endothelial cells located in a structure called the yolk sac that surrounds the embryo. Using a genetic-engineering approach to produce mouse embryos in which yolk-sac-derived EMPs and all their descendants were labelled with a fluorescent protein, Plein and colleagues unexpectedly found that these cells also contribute to the walls of blood vessels.

Analysis of the labelled cells revealed that

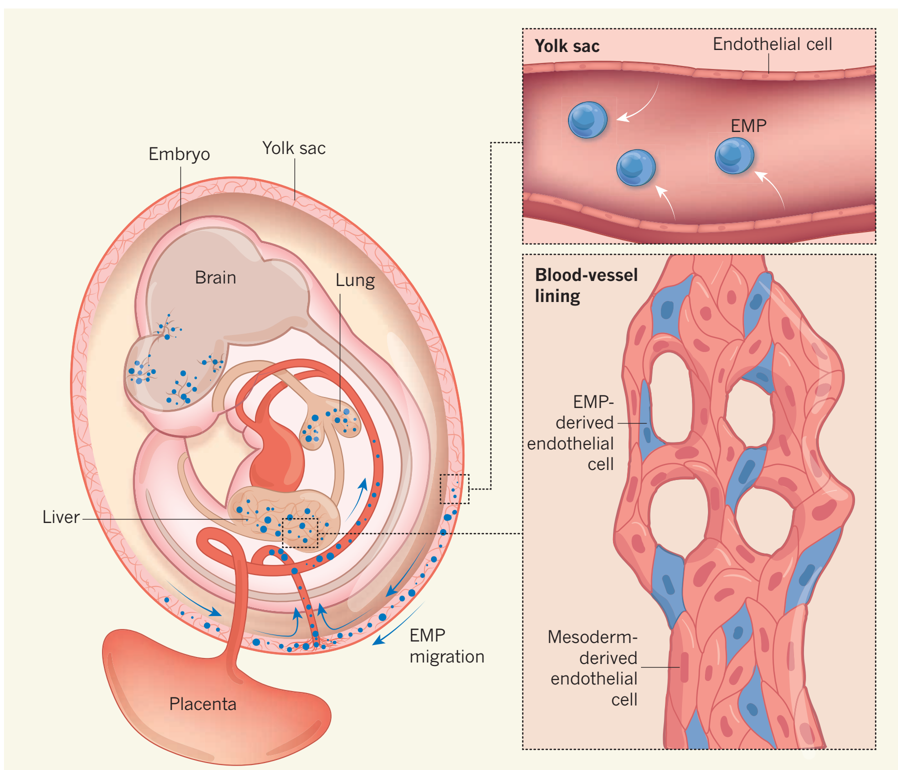


Figure 1 | Two contributors to the blood-vessel lining. An embryonic tissue called mesoderm (not shown) gives rise to endothelial cells, which proliferate to form both the inner lining of blood vessels and the lining of a structure called the yolk sac that surrounds developing embryos. Endothelial cells of the yolk sac in turn give rise (white arrows) to cells called erythro-myeloid progenitors (EMPs), which migrate into the embryo and are known to differentiate into embryonic blood-cell lineages. Plein *et al.*[3] demonstrate in mice that migrating EMPs can also revert to an endothelial-cell type. EMP-derived endothelial cells are incorporated into mesoderm-derived blood vessels in developing organs such as the brain, liver and lung, forming a mosaic pattern across the vessel lining.

50 Years Ago

A grant of $400,000 has been awarded to the University of Alberta by the National Research Council of Canada for the construction of a "controlled environment greenhouse" in which plants and animals native to the northern areas of Canada can be studied. The greenhouse, which is the first of its kind in Canada, will be one of several controlled environment facilities to be built for the university's department of botany at a total cost in excess of $1 million … Extending over 1,384 square feet, the greenhouse will contain several rooms in which different northern and mountainous environments can be simulated, so that long-term ecological and physiological studies of arctic, boreal and alpine plants can be carried out.
From *Nature* 12 October 1968

100 Years Ago

Rather more than four years ago an American metallurgist, in opening a discussion on the metallurgy of zinc, said wittily: "It is a time-honoured custom to throw bricks at the zinc man. The accusation is that he has borrowed a lime kiln and a gas retort and part of a sulphuric acid plant, hitched them together, and spent the last fifty years in regarding with holy veneration the reactions which take place in that retort. The copper man who thinks of zinc as something with which copper is adulterated to make brass, and the iron man who regards it as a sort of paint for corrugated sheets, and the lead man whose opinion as to zinc is not fit for publication, have long felt that when two or three of the minor details of their respective metallurgies were put in order, they would take a few days and fix up zinc on a modern basis."
From *Nature* 10 October 1918

EMPs actively migrate from the yolk sac into the embryo and differentiate into endothelial cells — reverting to their initial endothelial fate but now in an intraembryonic site. Unlike mesoderm-derived endothelial cells, which form blood vessels through local proliferation, the authors found that EMP-derived endothelial cells contribute to the vasculature of several organs by becoming incorporated into existing vessels and being interspersed in the mesoderm-derived endothelium, where they remain into adulthood (Fig. 1).

In 2015, the same genetic strategy was used to show[4] that adult immune cells called tissue-resident macrophages are derived from yolk-sac EMPs. This result surprised researchers in the field — until then, it had been thought that macrophages differentiated only from circulating white blood cells called monocytes. Thus, this EMP population constitutes a versatile group of cells. It has the potential to generate the primitive red blood cells and immune cells needed transiently during embryonic life, but can also generate tissue-resident macrophages and endothelial cells whose progeny persist in adults.

Plein *et al.* found that the percentage of endothelial cells in adult blood vessels that originated from EMPs ranged from about 30% in the brain to 60% in the liver. They showed that EMP-derived endothelial cells expressed high levels of the gene *Hoxa*, and that loss of *Hoxa* expression altered vessel development in the brain. Loss of *Hoxa* also affected brain-specific immune cells called microglia, making it hard to say for certain that the defects were caused solely by changes in EMP-derived endothelial cells. Nonetheless, these findings suggest an essential developmental requirement for EMP-derived endothelium in the brain.

The authors also examined the gene-expression profiles of endothelial cells in blood vessels. They found that the EMP-derived cells had a transcriptional signature consistent with the complete acquisition of an endothelial fate. However, there were some slight differences between these cells and neighbours of direct mesodermal descent. For example, the authors found over-representation of genes characteristic of a type of liver vessel in EMP-derived cells, and a lower representation of brain-specific markers of endothelial cells.

Taken together, Plein and colleagues' experiments showed that the vasculature of the embryo expands from two distinct lineages. Why does this matter? The origins of these cells are not only of intellectual interest, but could also have implications for physiology and disease. Although only speculation at this point, it is conceivable that endothelial cells from different developmental origins respond differently to the same stressor, as has been found for other lineages.

For example, vascular smooth-muscle cells, which form contractile muscle layers under the endothelium, originate from three distinct embryonic sources[5]. The sources affect the cells' gene-expression profiles and responses to pathological states[6]. They are also thought to be the reason that different regions of the vasculature react differently when exposed to the same stimulus. Following kidney failure in mice, patterns of vessel calcification differ in regions of the aorta (the body's largest blood vessel) that have distinct embryonic origins[7]. Mutations in a gene called *NT5E* in people result in vascular calcification exclusively in the limbs[8]. Finally, aneurysms, in which the blood-vessel wall weakens and bulges, seem to be triggered by different stressors in regions of blood vessels that have distinct origins[9].

Could distinct lineage histories also cause differential endothelial-cell responses to stimuli? This remains an open question, but the idea raises the possibility that the endothelium responds as a functional mosaic. Whereas large sections of vascular smooth muscle are derived from the same developmental source, it seems that EMP-derived endothelial cells interlace with cells of direct mesodermal origin. As such, alternative responses to stimuli might occur in the same segment of endothelium.

Interestingly, the endothelial lining of the aorta houses cells that have different proliferative abilities — cells capable of regenerating adult vessels exist side by side with cells that have a lower proliferative potential[10]. Perhaps this variability relates to the origin of these cells. Extending this idea, maybe the high percentage of EMP-derived endothelial cells in the liver is a factor in that organ's remarkable capacity for regeneration. Plein and colleagues' work will most certainly inspire investigators to pursue new experiments that explore the relationship between the origin of endothelial cells and their function.

Going forward, the degree to which these findings apply to humans needs to be formally tested. Naturally, lineage tracing is not feasible in humans. An alternative strategy would be to identify evolutionarily conserved gene-expression patterns characteristic of the two types of endothelial-cell lineage in mice, and to search for cells that have each profile in humans. It would also be exciting to clarify whether these two lineages differentially contribute to vessel repair following damage. ■

M. Luisa Iruela-Arispe *is in the Department of Molecular, Cell and Developmental Biology, and at the Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California 90095, USA.*
e-mail: arispe@mcdb.ucla.edu

1. Zovein, A. C. *et al. Cell Stem Cell* **3**, 625–636 (2008).
2. Gritz, E. & Hirschi, K. K. *Cell. Mol. Life Sci.* **73**, 1547–1567 (2016).
3. Plein, A., Fantin, A., Denti, L., Pollard, J. W. & Ruhrberg, C. *Nature* **562**, 223–228 (2018).
4. Gomez Perdiguero, E. *et al. Nature* **518**, 547–551 (2015).
5. Majesky, M. W. *Arterioscler. Thromb. Vasc. Biol.* **27**, 1248–1258 (2007).
6. Cheng, C., Bernardo, A. S., Trotter, M. W. B., Pedersen, R. A. & Sinha, S. *Nature Biotechnol.* **30**,

165–173 (2012).
7. Leroux-Berger, M. *et al. J. Bone Miner. Res.* **26**, 1543–1553 (2011).
8. St Hilaire, C. *et al. N. Engl. J. Med.* **364**, 432–442 (2011).
9. Lindsay, M. E. & Dietz, H. C. *Nature* **473**, 308–316 (2011).
10. McDonald, A. I. *Cell Stem Cell* **23**, 210–225 (2018).

This article was published online on 26 September 2018.

OPTOELECTRONICS

LED technology breaks performance barrier

Light-emitting diodes made from perovskite semiconductors have reached a milestone in the efficiency with which they emit light — potentially ushering in a new platform for lighting and display technology. SEE LETTERS P.245 & P.249

PAUL MEREDITH & ARDALAN ARMIN

Light-emitting diodes (LEDs) have revolutionized lighting and displays, not least because they use energy more efficiently than any previous light-emitting technology. Micro-LEDs made from inorganic, 'compound' semiconductors are emerging that deliver unprecedented resolution for displays, whereas organic semiconductor LEDs (OLEDs) provide unparalleled colour quality and near-180° viewing angles, and could potentially be used to develop flexible, lightweight displays. In this issue of *Nature*, two papers[1,2] report what could be the birth of a new family of LEDs based on semiconductors called perovskites. Remarkably, the efficiencies with which the perovskite LEDs (PLEDs) produce light from electrons already rival those of the best-performing OLEDs[3], and have been achieved in less than four years since the report[4] of the first PLED — suggesting that there is plenty of room for even further improvement in their performance.

Perovskites have shot to scientific stardom in the past few years, mostly because they show great promise for solar cells[5], but their potential for use in other applications, such as light sensors[6] and LEDs[4], is rapidly emerging. Crucially, perovskites can be processed from solution (for example, using low-cost, low-tech printing methods), and work well in the designs for optoelectronic devices that are easiest to make. This might allow perovskite-based devices that have large areas (several square centimetres) to be made extremely cheaply, and with low embodied energy (the total energy involved in the entire life cycle of a device).

Cao *et al.*[1] (page 249) and Lin *et al.*[2] (page 245) have independently developed PLEDs that break an important technological barrier: the external quantum efficiency (EQE) of the devices, which quantifies the number of photons produced per electron consumed, is greater than 20%. There are several similarities between the devices reported by the two groups. Perhaps most notably, the active (emissive) perovskite layer is about 200 nanometres thick in both cases, and is sandwiched between two relatively simple electrodes. This design is called a planar structure, and is the most basic manifestation of diodes made from thin films of materials (Fig. 1). The electrodes are appropriately modified to ensure that electrons and holes (quasiparticles formed by the absence of electrons in atomic lattices) are efficiently pumped into the perovskite. As in all LEDs, when electrons meet holes, they can release energy in the form of photons through a process known as radiative recombination.

Another similarity between the devices is that the perovskite layers were prepared using solutions, from which the semiconductors crystallized to form the emissive components of the LEDs. Cao *et al.* used a perovskite known as formamidinium lead iodide (FAPI), mixed with an amino-acid additive (aminovaleric acid) to control the size and orientation of the resultant perovskite crystals. FAPI has been quite widely explored as a semiconductor for solar cells, but Lin *et al.* report a new composite material in which crystals of the perovskite $CsPbBr_3$ (Cs, caesium; Pb, lead; Br, bromine) are partly enclosed by a shell of an organic compound (methyl ammonium bromide; MABr).

Achieving high EQEs in any LED requires the elimination of non-radiative losses — electron–hole-recombination pathways that do not produce photons. Both Cao and colleagues' and Lin and colleagues' PLEDs deliver on this equally well. But the two groups also used other, subtly different methods to improve the EQE.

Cao *et al.* targeted the outcoupling problem, which is well known to those working with thin-film LEDs (such as PLEDs and OLEDs). The outcoupling problem is that the optical physics of planar diodes causes 70–80% of the light generated by the semiconductor to be

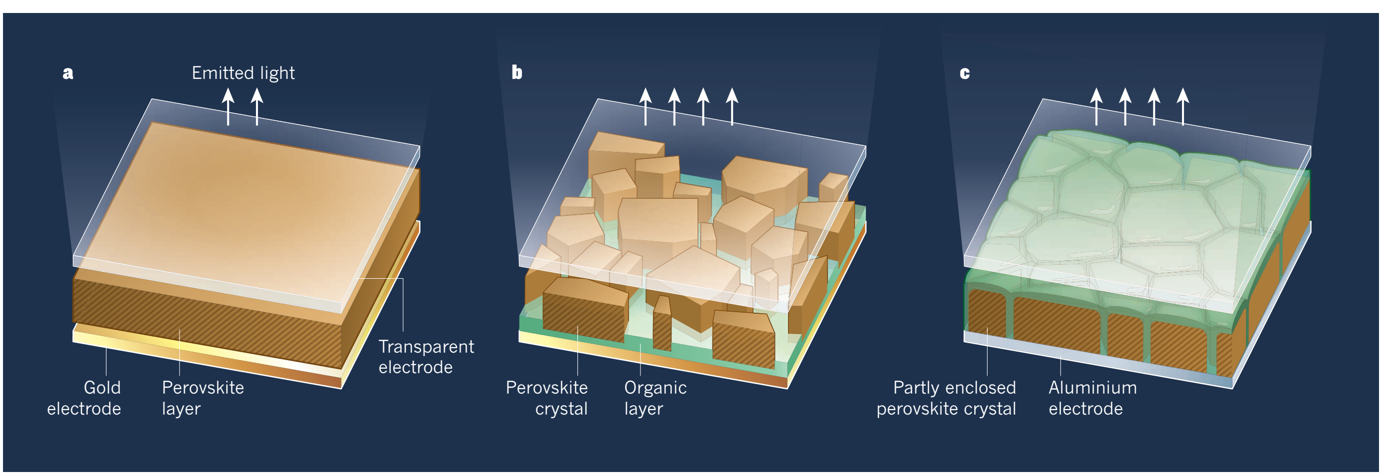


Figure 1 | Improved light-emitting diodes (LEDs) based on perovskite semiconductors. **a**, LEDs have previously been made from perovskites by sandwiching a thin layer of the semiconductor between a gold electrode and a transparent electrode. However, only about 20% of the light generated in the perovskite escapes from the device. **b**, Cao *et al.*[1] report perovskite LEDs (PLEDs) in which the semiconductor layer consists of separated submicrometre-sized crystals, partitioned from the gold electrode by a thin layer of an organic material. This design increases the amount of light that escapes. **c**, Lin *et al.*[2] report PLEDs based on a different perovskite, in which the semiconductor crystals are partly enclosed by an organic compound and the gold electrode is replaced by an aluminium one. This device optimizes the efficiency with which charges (not shown) that are pumped into the perovskite are converted into photons.

trapped in the device. Various strategies have attempted to address this issue in OLEDs, such as using diffraction gratings[7] and buckling the device[8].

But Cao and colleagues took a simpler approach: they optimized their perovskite-processing conditions so that the emissive layer spontaneously forms as distinct sub-micrometre-scale crystal platelets (Fig. 1). The authors' computational modelling shows that this submicrometre structuring increases the fraction of light that makes it out of the emissive layer to 30%, compared with 22% for an equivalent 'flat-layer' perovskite device (a device in which the perovskite layer does not have submicrometre structuring). In combination with the reduction in non-radiative losses, this results in an EQE of 20.7%.

By contrast, Lin *et al.* used a flat emissive layer, but tried to optimize the balance of electrons and holes injected into the perovskite, to make the most efficient use of every charge. This seems to be facilitated by the MABr shells that enclose the perovskite crystals. The resulting PLEDs have an EQE of 20.3%.

But caution is advised before ordering your PLED ultrahigh-definition television. OLEDs, and indeed all optoelectronic devices based on organic semiconductors, suffered for many years from stability issues. The first polymer OLEDs[9] could emit light for only seconds, and subsequent advances were needed to ensure that smartphone screens and OLED televisions last for tens of thousands of hours. The lifetime of LEDs can be measured by the T_{50} metric, which is the time for the performance of the device to drop by half. The T_{50} values of Cao and colleagues' and Lin and colleagues' PLEDs are currently modest: 20 hours and 100 hours, respectively.

Furthermore, displays require a minimum of three colours (and preferably more) to create high-quality colour images. Developing a range of colours for OLEDs was a big challenge. Cao and co-workers' PLED emits in the near-infrared region of the electromagnetic spectrum, and Lin and co-workers' PLED emits green light — which is definitely a good start. Multiple colours of PLEDs could be generated by altering the composition of the devices, but the same developmental journey as was needed for OLEDs lies ahead.

The two papers also highlight problems that occur every time new optoelectronic materials emerge as a technological platform: inconsistent characterization and a lack of standards. Because Cao and colleagues' PLED emits light from outside the visible spectrum, they report the metrics of their devices radiometrically — they use a measure that simply takes into account the total emitted power. By contrast, Lin and colleagues describe the emission of their green PLED using photometric measures, which are weighted by the response of the human eye. The two groups also report the peak EQEs at different brightnesses, and therefore at different driving currents. This makes direct comparison somewhat problematic.

Caveats aside, the two papers are a milestone in PLED development. For now, LEDs based on compound semiconductors remain the dominant technology: they outclass the competition in many respects, including cost, efficiency, colour and brightness. They will be hard to beat. But that should not stop the pioneers of perovskite (or, indeed, organic) LEDs from trying. ■

Paul Meredith *and* **Ardalan Armin** *are in the Department of Physics, Swansea University, Singleton Park, Swansea SA2 8PP, UK.*
e-mails: paul.meredith@swansea.ac.uk; ardalan.armin@swansea.ac.uk

1. Cao, Y. *et al. Nature* **562**, 249–253 (2018).
2. Lin, K. *et al. Nature* **562**, 245–248 (2018).
3. Di, D. *et al. Science* **356**, 159–163 (2017).
4. Tan, Z.-K. *et al. Nature Nanotechnol.* **9**, 687–692 (2014).
5. Lee, M. M., Teuscher, J., Miyasaka, T., Murakami, T. N. & Snaith, H. J. *Science* **338**, 643–647 (2012).
6. Lin, Q., Armin, A., Burn, P. L. & Meredith, P. *Nature Photon.* **9**, 687–694 (2015).
7. Ziebarth, J. M., Saafir, A. K., Fan, S. & McGehee, M. D. *Adv. Funct. Mater.* **14**, 451–456 (2004).
8. Koo, W. H. *et al. Nature Photon.* **4**, 222–226 (2010).
9. Burroughes, J. H. *et al. Nature* **347**, 539–541 (1990).

ANIMAL BEHAVIOUR

Foraging skills develop over generations

The movements of relocated wild animals reveal that a lost migratory skill was regained over successive generations. This suggests that skill improvements can occur over time as animals learn expertise from each other.

ANDREW WHITEN

The transmission of behavioural traditions by learning from others — cultural learning — was once thought to be a uniquely human attribute. However, evidence increasingly indicates that this phenomenon is widespread among animals, shaping behaviours from foraging for food to mate choice to predator avoidance[1]. Claims for human uniqueness in our cultural skills have therefore been pinned on our species' capacity for what is called cumulative culture: the ramping up of cultural sophistication as each generation builds on their ancestors' cultural achievements[2,3]. Writing in *Science*, Jesmer *et al.*[4] now challenge this view in a study of the development of migratory skill in wild populations of bighorn sheep (*Ovis canadensis*) and moose (*Alces alces*) populations that have been moved to unfamiliar locations. Their findings have implications for understanding the evolution of cumulative culture in both humans and other animals, and for conservation policies[5,6].

In the wild, bighorn sheep (Fig. 1) and moose normally migrate in spring and move between distinct seasonal ranges. These movements follow a pattern known as green-wave surfing, whereby the animals' migration tracks the availability of high-quality vegetation, which peaks at different times in different places depending on factors such as altitude. How animals evolved the capacity for this type of migratory behaviour remains unknown.

Jesmer and colleagues investigated the migration of bighorn sheep and moose that had been moved to unfamiliar areas in recent decades to repopulate regions in which these types of animal had been wiped out by disease or hunting. The authors compared the migration of such relocated populations with that of animals in long-established populations that had been migrating for many generations in a particular region. They noted that when individuals had been moved to an unfamiliar location, the animals usually ceased migrating, although migratory behaviour gradually re-emerged in subsequent generations.

The researchers fitted animals with a collar containing a Global Positioning System (GPS) device that enables accurate tracking of an animal's position. They combined this information with the corresponding satellite imagery for the region that revealed where and when vegetation was at peak quality. To measure animals' green-wave surfing skills, the authors counted the number of days between the peak forage quality at a location and the arrival of an animal there. When the authors analysed bighorn sheep from migratory populations that had been relocated at times ranging from 0 to 35 years ago, these animals surfed the green wave approximately half as effectively as animals from populations that had been established in a particular region for more than 200 years.

Jesmer and colleagues then combined these and other bighorn records with similar data for moose that had been relocated to a given region between 10 and 110 years ago. The combined results for these 267 bighorn and 189 moose were consistent with a model in which it took up to 30 years (between 4 and 5 generations) for migration to distinct

DONALD M. JONES/MINDEN PICTURES/FLPA

Figure 1 | Bighorn sheep *(Ovis canadensis)* in Montana. Jesmer *et al.*[4] report that the migration of wild bighorn sheep so as to track the spring arrival of high-quality vegetation is a skill that develops over many generations.

seasonal ranges to re-emerge in these species. It took almost a century for a relocated population to reach a point at which half its number migrated in this way. Animals that do not migrate to distinct seasonal ranges might begin to undertake green-wave surfing over small distances.

The bighorn-sheep data span more than two centuries, and the authors found that migratory behaviour had spread to nearly all of the bighorn sheep individuals that had been established in a location for at least around 200 years. Most interestingly, green-wave-surfing knowledge steadily increased over the decades, indicating that migratory skill progressively rises to the highest levels over long time frames.

The authors suggest that their findings can be explained by a cumulative process of acquisition of migratory skill involving cycles of individual and cultural learning that span many decades and generations. Individuals might acquire some initial surfing knowledge by personal learning, which then becomes available to their young through social learning, and the next generation might build on this knowledge through further exploration. The refinement of skills in the next generation could be similarly enhanced, and so on. Repeated cycles of individual and social learning might thus generate a cumulative culture of progressively refined surfing expertise and an increase in the proportion of migrants in the population.

Unfortunately, no direct evidence of social learning in these animals has yet been documented in the wild to support this interesting idea. However, a previous analysis of the homing of domesticated pigeons[7] provides data suggesting that cumulative effects of social learning can occur in animals. In this study, two birds were tracked using GPS monitoring as they flew homing flights together. One animal of the pair was then replaced by a pigeon that had not flown the route before, and this pair of birds flew a series of homing flights. After a series of successive replacements of one bird of the pair, the efficiency of the homing flights improved significantly from that observed at the outset. The birds in the later pairings were different from those that made the initial flights, so this improvement is consistent with a model of individual learning coupled with social transmission across these 'cultural generations'.

"Migratory skill progressively rises to the highest levels over long time frames."

The bighorn and moose findings might well reflect similar learning processes. Moreover, for these animals, cultural learning will probably involve the acquisition of a diverse range of expertise relating to different aspects of migration in addition to green-wave surfing skill, such as knowledge of the predation risks in what are known as 'landscapes of fear'[8], which is of particular consequence given that offspring migrate with their mothers. The findings of Jesmer *et al.* provide an advance for this area of research by investigating learning in the wild, across multiple generations and over many decades, illuminating our understanding of animal culture and the collective behaviour of a population over time.

Jesmer and colleagues interpret the long-term growth over time in the populations' green-wave surfing skills to imply that, over successive generations, the individuals of a particular population develop more-refined migration skills than those in earlier generations. However, a possibility worth investigating is whether the improvements in a relocated population's ability to track peak vegetation might be driven mainly by an increase in the proportion of animals that learn migratory skills from others, rather than because the migratory skills of individuals increase over successive generations. Nevertheless, what would develop under this scenario is also the progressive, collective enhancement in migration skills of the population as a whole, an example that is relevant to the topic of collective intelligence in animal groups[9,10].

The United Nations Environment Programme has been considering how evidence for cultural learning in animals should inform conservation policies. This is of particular note for animal populations that migrate through, or are located in, areas that cross national borders. A panel of scientists has recently assembled key evidence and recommendations related to this in a report for policymakers[6]. The findings of Jesmer *et al.* underscore the importance of such considerations if wild-animal populations develop skills that enhance their survival over a time span of centuries. In the case of migratory skills, the blocking of traditional migratory routes by human-made barriers such as roads could lead to the loss of animals' hard-won cultural knowledge.

Conservation efforts need to take into account the significance of such knowledge, the scope of which we are perhaps only starting to recognize[1,10], and our understanding of which is extended by long-term perspectives such as those reported by Jesmer and colleagues. Cumulative culture of this kind might be more widespread in nature than was previously assumed, and not unique to humans. Accordingly, understanding the gulf between these and our own species' cumulative cultures might require us to consider more-specific aspects of cultural transmission, including modes of learning such as intentional teaching, or cultural contents, such as adopting qualitatively improved materials for tools. As the latter example suggests, human culture could progress by incorporating qualitatively distinct innovations. It remains a controversial question whether this ability is also found in animals — can they go beyond just achieving gradual refinements in a skill, such as green-wave surfing, to add a transformative new approach to solve a particular problem? ■

Andrew Whiten *is in the Centre for Social Learning and Cognitive Evolution, School of Psychology and Neuroscience, University of St Andrews, St Andrews KY16 9JP, UK.*
e-mail: awhiten@st-andrews.ac.uk

1. Whiten, A., Ayala, F. J., Feldman, M. W. & Laland, K. N. *Proc. Natl Acad. Sci. USA* **114**, 7775–7781 (2017).
2. Tennie, C., Call, J. & Tomasello, M. *Phil. Trans. R. Soc. B* **364**, 2405–2415 (2009).
3. Boyd, R., Richerson, P. J. & Henrich, J. *Proc. Natl Acad. Sci. USA* **108** (Suppl. 2), 10918–10925 (2011).
4. Jesmer, B. R. *et al. Science* **361**, 1023–1025 (2018).
5. Festa-Bianchet, M. *Science* **361**, 972–973 (2018).
6. *Report of the CMS Workshop on Conservation Implications of Animal Culture and Social Complexity* Available at go.nature.com/2qc6esk (2018).
7. Sasaki, T. & Biro, D. *Nature Commun.* **8**, 15049 (2017).
8. Laundré, J. W., Hernández, L. & Alterndorf, K. B. *Can. J. Zool.* **79**, 1401–1409 (2001).
9. Biro, D., Sasaki, T. & Portugal, S. J. *Trends Ecol. Evol.* **31**, 550–562 (2016).
10. Berdahl, A. M. *et al. Phil. Trans. R. Soc. B* **373**, 20170009 (2018).

This article was published online on 1 October 2018.

QUANTUM PHYSICS

Unexpected noise from hot electrons

Experiments reveal a previously unreported type of electronic noise that is caused by temperature gradients. The finding has practical implications, and could help in detecting unwanted hotspots in electrical circuits. SEE LETTER P.240

ELKE SCHEER & WOLFGANG BELZIG

A fundamental feature of any electrical measurement is noise — random and uncorrelated fluctuations of signals. Although noise is typically regarded as undesirable, it can be used to probe quantum effects and thermodynamic quantities. On page 240, Shein Lumbroso *et al.*[1] report the discovery of a type of electronic noise that is distinct from all others previously observed. Understanding such noise could be essential for designing efficient nanoscale electronics.

A century ago, the German physicist Walter Schottky published a seminal paper that described different causes and manifestations of noise in electrical measurements[2]. Schottky showed that an electric current produced by an applied voltage is noisy, even at absolute zero temperature, when all random heat-induced motion has stopped. This noise is a direct consequence of the fact that electric charge is quantized — it comes in discrete units. Because the noise results from the granularity of the charge flow, it is called shot noise.

It was already known at the time of Schottky's work that, in systems that are in thermal equilibrium, noise with distinctly different properties from shot noise comes into play at non-zero temperatures — this is known as thermal (Johnson–Nyquist) noise. Today, shot noise is a key tool for characterizing nanoscale electrical conductors, because it contains information about quantum-transport properties that cannot be revealed from mere electric-current measurements[3,4].

Shein Lumbroso *et al.* studied junctions composed of single atoms or molecules suspended between a pair of gold electrodes. The authors fabricated the electrodes by breaking a thin gold wire into two parts and bringing the parts gently back into contact. They evaporated hydrogen molecules on to this device, which is known as a mechanically controllable

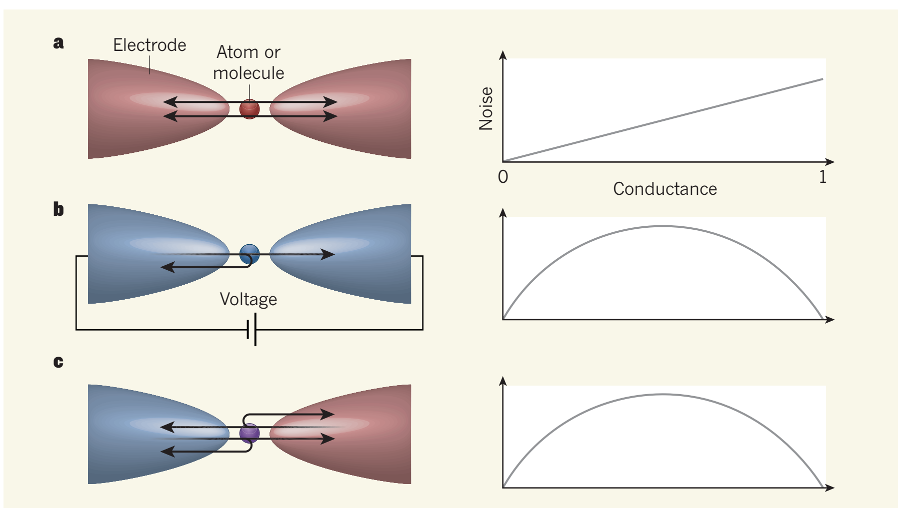


Figure 1 | Three types of electronic noise. Shein Lumbroso *et al.*[1] report experiments in which single atoms or molecules are suspended between the tips of two electrodes. **a**, At a non-zero temperature (red), electrons flow between the two electrodes (arrows). The electrical signal associated with this motion contains a type of noise called thermal noise, which varies linearly with electrical conductance (shown here in units of the quantum of conductance). **b**, If a voltage is applied to the device, electrons flow from one electrode to the other, and can be backscattered from the atom or molecule. The resulting signal contains 'shot' noise that is present even when the device is at absolute zero temperature (blue). Shot noise has a characteristic (non-monotonic) dependence on conductance. **c**, If a temperature gradient is applied to the device (indicated by rising temperatures from blue to purple to red), electrons flow from both of the electrodes and can be backscattered. The authors show that the resulting electrical signal contains a previously unreported type of noise, which they term delta-T noise. This noise has the same dependence on conductance as shot noise.

break junction, so that individual hydrogen atoms or molecules were captured between the tips of the electrodes, thereby establishing an electrical contact.

The resulting junctions constituted a single quantum-mechanical transport channel in which electrons could be transmitted from one electrode to the other with a probability that could be adjusted by varying the openness of the channel. This set-up provided an ideal test bed for exploring the properties of the so-far-overlooked noise contribution.

The authors observed a strong increase in electronic noise when they applied a temperature difference between the two electrodes, compared with when the electrodes were at the same temperature. The additional noise, which the authors call delta-T noise, scaled with the square of the temperature difference. It exhibited the same dependence on electrical conductance as shot noise (Fig. 1).

Shein Lumbroso and colleagues explained their finding using the quantum theory of charge transport, known as the Landauer theory[5], which has been developed in the past few decades. This theory incorporates both shot noise and thermal noise, and has been tested intensively down to the atomic and molecular scale[3]. It has been found to accurately describe many experimental observations obtained when working entirely in thermal equilibrium, or when applying small voltages. The authors took a closer look at the theory, and found that it includes a noise component that occurs when solely a temperature difference is applied across a junction: delta-T noise.

It is well established that an electric current can arise from a temperature difference in the absence of an applied voltage — a phenomenon called the Seebeck effect. However, delta-T noise is not the shot noise associated with this thermally induced current. The authors' results indicate that delta-T noise is larger than this shot noise, and has a different dependence on the temperature difference. Instead, the results suggest that delta-T noise arises from the discreteness of the charge carriers mediating the heat transport.

Because the Landauer theory is widely used, it is surprising that delta-T noise has not previously been observed. The importance of carefully considering all of the spatial temperature differences and resulting electric currents to understand the current flow in atomic and molecular contacts was pointed out in a 2013 paper[6], but implications for noise were not addressed.

Shein Lumbroso *et al.* found that the Landauer theory accurately describes all of the characteristic properties of delta-T noise. In this sense, their experiments are yet another beautiful demonstration of the theory. But the work also conveys a key message: careful design and rigorous analysis of experiments are required when studying any of the details of quantum transport.

The authors' discovery also has practical implications. In particular, quantum-transport experiments that are not entirely in thermal equilibrium could show strongly enhanced noise, which might be mistaken for noise arising from interactions between the charge carriers or from other subtle effects. Experimentalists who wonder about finding unexpectedly high noise in their electric-current measurements might wish to revisit their set-ups to search for unintentional temperature gradients. The most practical application of the authors' work is probably that the enhanced noise could be used to detect unwanted hotspots in electrical circuits.

For the future, researchers could explore the relationship between delta-T noise and shot noise that has a nonlinear dependence on applied voltage, which was observed earlier this year in high-voltage experiments on atomic junctions[7]. Such studies could also be expanded to more-complex quantum-transport experiments — for instance, those on artificial atoms called quantum dots. Because of the sensitivity of delta-T noise to the properties and interactions of charge carriers, the phenomenon might become a valuable tool in quantum-transport investigations. ■

Elke Scheer *and* **Wolfgang Belzig** *are in the Department of Physics, University of Konstanz, 78464 Konstanz, Germany.*
e-mails: elke.scheer@uni-konstanz.de; wolfgang.belzig@uni-konstanz.de

1. Shein Lumbroso, O., Simine, L., Nitzan, A., Segal, D. & Tal, O. *Nature* **562**, 240–244 (2018).
2. Schottky, W. *Ann. Phys.* **362**, 541–567 (1918).
3. Blanter, Y. M. & Büttiker, M. *Phys. Rep.* **336**, 1–166 (2000).
4. Beenakker, C. & Schönenberger, C. *Phys. Today* **56**, 37–42 (2003).
5. Landauer, R. *IBM J. Res. Dev.* **1**, 223–231 (1957).
6. Lee, W. *et al. Nature* **498**, 209–212 (2013).
7. Tewari, S. & van Ruitenbeek, J. *Nano Lett.* **18**, 5217–5223 (2018).

Competing interests declared: see go.nature.com/2zqqjdv for details.

CANCER

Thousands of short cuts to genetic testing

Gene editing has now been used to introduce every possible single-nucleotide mutation into key protein-coding regions in the cancer-predisposition gene *BRCA1*, to identify the variants that are linked to cancer risk. See Article p.217

STEPHEN J. CHANOCK

For decades, cancer geneticists have been trying to understand which changes in the sequence of the *BRCA1* gene predispose affected individuals to developing breast or ovarian cancer. Extensive efforts have focused on interpreting the plethora of genetic variants in *BRCA1*, using clinical observations to determine whether this or that variant warrants patient counselling about options for medical intervention[1]. Generally, *BRCA1* variants are sorted into three categories[2,3]: benign variants, which cause no concern; deleterious variants, which can confer a high risk of cancer; and an unsettling intermediate known as variants of uncertain significance (VUS). Hardest to classify are variants that arise only rarely, of which there are thousands for *BRCA1*. Conventionally, genetic sleuthing has focused on families or populations within which certain mutations occurred at an unusually high frequency, exposing the effects of deleterious variants. But on page 217, Findlay *et al.*[4] report an innovative laboratory-based approach to assessing the effect of thousands of variants across protein-coding regions of *BRCA1*.

The BRCA1 protein is a key tumour suppressor, and is essential for a DNA-repair pathway called homology-directed repair. Mutations that prevent this function lead to the death of cultured human cells of a strain called HAP1 (ref. 5). Findlay and colleagues made clever use of this property of HAP1 cells to screen for deleterious *BRCA1* variants.

The authors used a gene-editing approach called CRISPR–Cas9 to accurately mutate each nucleotide in 13 crucial protein-coding regions (exons) of *BRCA1* into every other possible base, one nucleotide at a time — an exhaustive technique known as saturation genome editing (SGE). In each experiment, they edited 1 exon of *BRCA1* in 20 million HAP1 cells simultaneously. They left the cells to grow *in vitro* for 11 days, then sequenced the edited exon to gauge the frequency at which each variant was present in the cell population. From these data, they designated each variant as functional (if its frequency indicated that homology-directed repair was active in cells harbouring that variant), non-functional (if the frequency was lower than average, indicating that the variant led to cell death), or intermediate (Fig. 1).

Findlay *et al.* found that their results fit well with those obtained from a complementary assay designed to test whether homology-directed repair occurs normally in *BRCA1* mutant cells,

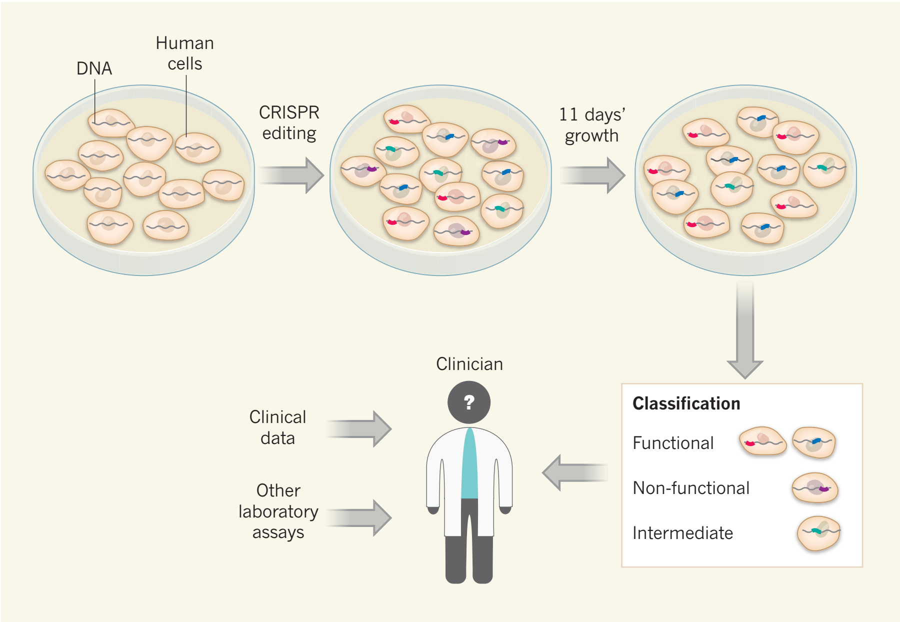


Figure 1 | Assaying how genetic variants affect BRCA1-protein function. Findlay *et al.*[4] grew human cells in culture. They used a gene-editing approach called CRISPR–Cas9 to modify the genomes of the cells in such a way that every possible single-nucleotide variation in a given protein-coding region (exon) of the *BRCA1* gene was present in some cells of the population. The edited cells were cultured for 11 days, and the exon was then sequenced (not shown) to determine the frequency of each variant in the population. Variants present at the expected frequency were classified as functional, meaning that the protein had no effect on BRCA1 function. Those present at lower-than-expected levels were designated as non-functional, because they had caused changes in BRCA1 that prevented normal cell growth. Variants in the middle of the range were designated as intermediate. This approach could be combined with other clinical data and with laboratory-based assays to enable accurate variant classification by clinicians, but how this should be done needs further discussion. (Adapted from Figure 1b of the paper[4].)

which is outlined in an accompanying paper in *The American Journal of Human Genetics*[6]. They also compared their results with an internationally recognized set of annotated *BRCA1* variants[7] designated as benign, deleterious or VUS on the basis of clinical data (or lack thereof, for many of the VUS). They found that their results, although not perfect, were strikingly accurate. Variants designated as non-functional in Findlay and colleagues' analysis generally corresponded with those annotated as deleterious in the database, and, reassuringly, nearly all functional variants corresponded with those annotated as benign.

The group therefore reasoned that its approach could be used to shed light on the many variants of the vexing VUS class, which keep clinicians up at night. The researchers provide evidence that some *BRCA1* VUS are non-functional — a subset that should be monitored carefully in the future. Finally, they provided insights into the extent to which variants in the sequences that flank exons can disrupt protein function, thus extending our ability to interpret more pieces of the genome.

The current study is remarkable for its scale, in that the method enables almost 4,000 possible *BRCA1* variants to be analysed in parallel. The next study should look at regions of *BRCA1* outside the 13 exons studied here, especially those that also harbour deleterious mutations and VUS. VUS are currently piling up, because the rate at which new patient sequences for *BRCA1* are being collected is greatly outstripping the accumulation of clinical information needed to classify variants. Findlay and colleagues' approach represents a potential game-changer for assessing VUS. But first, it will be crucial to collect further clinical data to validate the exciting findings of this paper.

If validated, the technique could prove to be a major advance over previous efforts to study the impact of VUS in the laboratory. Such efforts typically combined computational models with *in vitro* assays of, for example, protein–protein interactions or drug sensitivity. Over the past decade, these analyses have begun to be incorporated into annotation strategies. But the pace of change has been slow, and there is considerable disagreement over the weight that should be given to this type of evidence[8]. The scale of Findlay and colleagues' study, together with its apparent accuracy, bodes well for its future integration into the classification of *BRCA1* variants. It is likely that the findings will be incorporated into current efforts to annotate *BRCA1* variants[7] that are part of the international BRCA Challenge (http://brcaexchange.org).

> ***"It is likely that the findings will be incorporated into current efforts to annotate BRCA1 variants."***

But further thought is required to determine the best way to incorporate Findlay and colleagues' assay into variant classification. The backbone of genetic testing is the availability of sufficient clinical data to assign risk to a given variant[7,8]. The new assay should supplement, not supplant, these data. It might be tempting to make immediate use of the assay to interpret VUS identified during human genetic testing, particularly because SGE has been used successfully in the past to identify targets for drug development[9]. But *in vitro* data alone should not be used as the basis for medical advice — at least until the approach has been clinically validated.

Could Findlay and colleagues' approach be applied to analyse variants in the other 20,000 or so genes in the human genome? For cancer-predisposition genes (which number well over 100)[10], including the well-studied genes *BRCA2* and *TP53*, the answer is probably yes. For these genes, non-functional variants would be expected to alter cell growth in culture, enabling a modification of Findlay and colleagues' frequency assay to be used. The effort involved in developing such an assay for each gene is substantial, and will probably slow the immediate application of SGE for assessing VUS. But although developing these assays for all exons in cancer genes will take time and money, the dividends could be spectacular for cancer geneticists. Going forward, large SGE analyses of cancer genes should be made publicly available. It is plausible that SGE will lead to the identification of previously unknown cancer-predisposition genes that, in turn, astute clinicians will verify.

Findlay and co-workers' provocative paper should turn heads across disparate domains of genomics. It remains to be seen to what extent the authors' approach can be applied to all of these domains, or whether it will remain an exciting development restricted mainly to cancer. Either way, this study should help researchers to realize the promise of precision medicine[11]. ■

Stephen J. Chanock *is in the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.*
e-mail: chanocks@mail.nih.gov

1. Kuchenbaecker, K. B. *et al. J. Am. Med. Assoc.* **317**, 2402–2416 (2017).
2. Plon, S. E. *et al. Hum. Mutat.* **29**, 1282–1291 (2008).
3. Richards, S. *et al. Genet. Med.* **17**, 405–423 (2015).
4. Findlay, G. M. *et al. Nature* **562**, 217–222 (2018).
5. Blomen, V. A. *et al. Science* **350**, 1092–1096 (2015).
6. Starita, L. M. *et al. Am. J. Hum. Genet.* https://doi.org/10.1016/j.ajhg.2018.07.016 (2018).
7. Rehm, H. L. *et al. N. Engl. J. Med.* **372**, 2235–2242 (2015).
8. Spurdle, A. B. *et al. Hum. Mutat.* **33**, 2–7 (2012).
9. Ma, L. *et al. Proc. Natl Acad. Sci. USA* **114**, 11751–11756 (2017).
10. Rahman, N. *Nature* **505**, 302–308 (2014).
11. Collins, F. S. & Varmus, H. *N. Engl. J. Med.* **372**, 793–795 (2015).

This article was published online on 12 September 2018.

ARTICLE

OPEN

https://doi.org/10.1038/s41586-018-0579-z

The UK Biobank resource with deep phenotyping and genomic data

Clare Bycroft[1,13], Colin Freeman[1,13], Desislava Petkova[1,12,13], Gavin Band[1], Lloyd T. Elliott[2], Kevin Sharp[2], Allan Motyer[3], Damjan Vukcevic[3,4], Olivier Delaneau[5,6,7], Jared O'Connell[8], Adrian Cortes[1,9], Samantha Welsh[10], Alan Young[11], Mark Effingham[10], Gil McVean[1,11], Stephen Leslie[3,4], Naomi Allen[11], Peter Donnelly[1,2,14] & Jonathan Marchini[1,2,14]*

The UK Biobank project is a prospective cohort study with deep genetic and phenotypic data collected on approximately 500,000 individuals from across the United Kingdom, aged between 40 and 69 at recruitment. The open resource is unique in its size and scope. A rich variety of phenotypic and health-related information is available on each participant, including biological measurements, lifestyle indicators, biomarkers in blood and urine, and imaging of the body and brain. Follow-up information is provided by linking health and medical records. Genome-wide genotype data have been collected on all participants, providing many opportunities for the discovery of new genetic associations and the genetic bases of complex traits. Here we describe the centralized analysis of the genetic data, including genotype quality, properties of population structure and relatedness of the genetic data, and efficient phasing and genotype imputation that increases the number of testable variants to around 96 million. Classical allelic variation at 11 human leukocyte antigen genes was imputed, resulting in the recovery of signals with known associations between human leukocyte antigen alleles and many diseases.

Understanding the role that genetics has in phenotypic and disease variation, and its potential interactions with other factors, is crucial for a better understanding of human biology. It is hoped that this will lead to more successful drug development[1], and potentially to more efficient and personalized treatments. As such, a key component of the UK Biobank resource has been the collection of genome-wide genetic data on every participant using a purpose-designed genotyping array[2]. An interim release of genotype data on approximately 150,000 UK Biobank participants in May 2015[3] has already facilitated numerous studies[4–6].

In this paper, we summarize the existing and planned content of the phenotype resource and describe the genetic dataset on the full 500,000 participants. To facilitate its wider use, we applied a range of quality control procedures and conducted a set of analyses that reveal properties of the genetic data—such as population structure and relatedness—that can be important for downstream analyses. In addition, we estimated haplotypes and imputed genotypes into the dataset that increases the number of testable variants by more than 100-fold to approximately 96 million variants. We also imputed classical allelic variation at 11 human leukocyte antigen (HLA) genes, and replicated signals of known associations between HLA alleles and many common diseases. We describe tools that allow efficient genome-wide association studies (GWAS) of multiple traits and fast phenome-wide association studies, which work together with a new compressed file format that has been used to distribute the dataset. As a further check of the genotyped and imputed datasets, we performed a test-case genome-wide association scan on a well-studied human trait, standing height.

The UK Biobank

A wide variety of phenotypic information as well as biological samples have been collected for each of the approximately 500,000 UK Biobank participants (Fig. 1). At recruitment, participants provided electronic signed consent, answered questions on socio-demographic, lifestyle and health-related factors, and completed a range of physical measures (see Extended Data Table 1). They also provided blood, urine and saliva samples, which were stored in such a way as to allow many different types of assay to be performed (for example, genetic, proteomic and metabonomic analyses)[7]. Once recruitment was fully underway, further enhancements were introduced to the assessment visit, including a range of eye measures, an electrocardiograph test, arterial stiffness and a hearing test.

The baseline information has been, and will continue to be, extended in several ways. For example, repeat assessments are planned to be conducted in subsets of the cohort every few years, to enable calibration of measurements, adjustment for regression dilution, and estimation of longitudinal change. Objective measures of physical activity have also been collected (using a tri-axial accelerometer) in 100,000 participants in 2013–2014[8] with repeated measures being collected over a period of a year (on a seasonal basis) from 2,500 of these participants. A multi-modal imaging assessment is currently underway, which comprises magnetic resonance imaging (MRI) of the brain[9], heart[10] and body, carotid ultrasound[11] and a whole body dual-energy X-ray absorptiometry of the bones and joints[12]. Data collection started in 2014 and is anticipated to take 7–8 years to achieve imaging for 100,000 participants in dedicated imaging assessment centres across the United Kingdom, with repeat imaging measures being planned for a subset of participants.

All participants provided consent for follow-up through linkage to their health-related records. As of May 2018, there were over 14,000 deaths, 79,000 participants with cancer diagnoses, and

[1]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Department of Statistics, University of Oxford, Oxford, UK. [3]Melbourne Integrative Genomics and the Schools of Mathematics and Statistics, and BioSciences, The University of Melbourne, Parkville, Victoria, Australia. [4]Murdoch Children's Research Institute, Parkville, Victoria, Australia. [5]Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland. [6]Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland. [7]Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland. [8]Illumina Ltd, Chesterford Research Park, Little Chesterford, Essex, UK. [9]Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, John Radcliffe Hospital, University of Oxford, Oxford, UK. [10]UK Biobank, Adswood, Stockport, Cheshire, UK. [11]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [12]Present address: Procter & Gamble, Brussels, Belgium. [13]These authors contributed equally: Clare Bycroft, Colin Freeman, Desislava Petkova. [14]These authors jointly supervised this work: Peter Donnelly, Jonathan Marchini. *e-mail: marchini@stats.ox.ac.uk

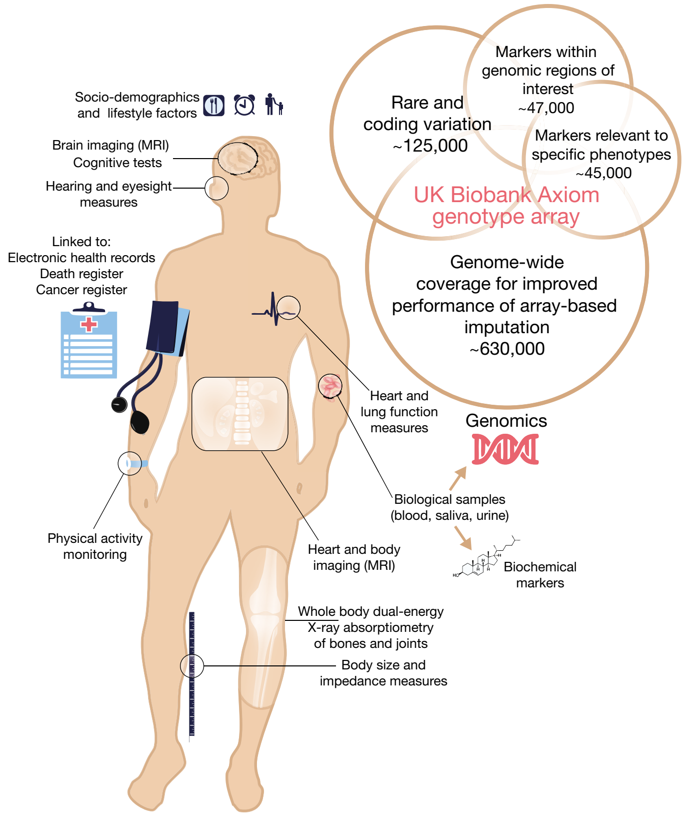


Fig. 1 | Summary of the UK Biobank resource and genotyping array content. Summary of the major components of the UK Biobank resource. See Extended Data Table 1 for more details. The figure also shows a schematic representation of the different categories of content on the UK Biobank Axiom genotype array. Numbers indicate the approximate count of markers within each category, ignoring any overlap. A more detailed description of the array content is available in the UK Biobank Axiom Array Content Summary[2].

400,000 participants with at least one hospital admission. Considerable efforts are now underway to incorporate data from a range of other national datasets including primary care, screening programmes, and disease-specific registries, as well as asking participants directly about health-related outcomes through online questionnaires (see Extended Data Table 1). Efforts are also underway to develop scalable approaches that can characterize in detail different health outcomes by cross-referencing multiple sources of coded clinical information[13].

Measurements for a wide range of biochemical markers of key interest to the research community have also been carried out, including those that have known associations with disease (for example, lipids for vascular disease and sex hormones for cancer), diagnostic value (for example, HbA_{1c} for diabetes and rheumatoid factor for arthritis), or the ability to characterize phenotypes not otherwise well assessed (for example, biomarkers for renal and liver function).

UK Biobank is an open-access resource that encourages researchers from around the world, including those from the academic, charity, public and commercial sectors, to access the data for any health-related research that is in the public interest.

Whole-genome genotyping

The UK Biobank genetic data contains genotypes for 488,377 participants. These were assayed using two very similar genotyping arrays. A subset of 49,950 participants involved in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study were genotyped at 807,411 markers using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix (now part of Thermo Fisher Scientific), which is described elsewhere[6]. Following this, 438,427 participants were genotyped using the closely related Applied Biosystems UK Biobank Axiom Array (825,927 markers) that shares 95% of marker content with the UK BiLEVE Axiom Array. The marker content of the UK Biobank Axiom array was chosen to capture genome-wide genetic variation (single nucleotide polymorphism (SNPs) and short insertions and deletions (indels)), and is summarized in Fig. 1. Many markers were included because of known associations with, or possible roles in, disease. The array also includes coding variants across a range of minor allele frequencies (MAFs), including rare markers (<1% MAF); and markers that provide good genome-wide coverage for imputation in European populations in the common (>5%) and low frequency (1–5%) MAF ranges. Further details of the array design are in the UK Biobank Axiom Array Content Summary[2].

DNA was extracted from stored blood samples that had been collected from participants on their visit to a UK Biobank assessment centre. Genotyping was carried out by Affymetrix Research Services Laboratory in 106 sequential batches of approximately 4,700 samples (see Methods, Supplementary Table 12). Affymetrix applied a custom genotype calling pipeline and quality filtering optimized for biobank-scale genotyping experiments and the novel genotyping arrays, which contain markers that had not been previously typed using Affymetrix technology (see Methods). This resulted in a set of genotype calls for 489,212 samples at 812,428 unique markers (biallelic SNPs and indels) from both arrays, with which we conducted further quality control and analysis (Extended Data Table 2).

Our quality control pipeline was designed specifically to accommodate the large-scale dataset of ethnically diverse participants, genotyped in many batches, using two slightly different arrays, and which will be used by many researchers to tackle a wide variety of research questions. Participants reported their ethnic background by selecting from a fixed set of categories[14]. Although most (94%) individuals report their ethnic background as within the broad-level group ‘white’, there are still approximately 22,000 individuals with a self-reported ethnic background originating outside Europe (Extended Data Table 3). We used approaches based on principal component analysis (PCA) to account

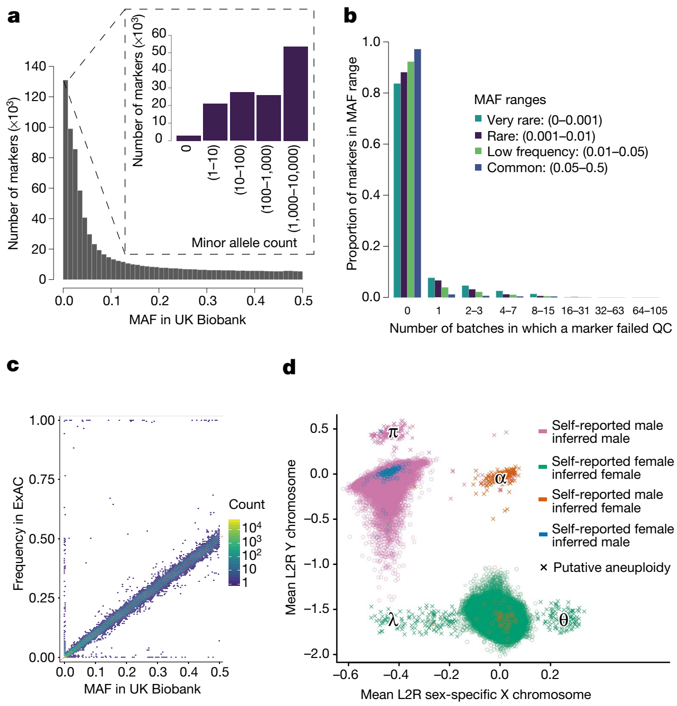


Fig. 2 | Summary of genotype data quality and content. All plots show properties of the UK Biobank genotype data after applying quality control. **a**, MAF distribution based on all samples (805,426 markers). The inset shows rare markers only (MAF < 0.01). **b**, The distribution of the number of batch-level quality control (QC) tests that a marker fails (see Methods). For each of four MAF ranges, we show the fraction of markers that fail the specified number of batches. **c**, Comparison of MAF in UK Biobank with the frequency of the same allele in ExAC, among the European-ancestry participants within each study (Supplementary Information). This analysis used 91,298 overlapping markers. Each hexagonal bin is coloured according to the number of markers falling in that bin ($\log_{10}$ scale). The dashed red line shows $x = y$. The markers with very different allele frequencies seen on the top, bottom and left-hand sides of the plot comprise approximately 300 markers. This is 0.3% of all markers in the comparison (see Supplementary Information for discussion). **d**, Mean $\log_2$ ratios (L_2R) on X and Y chromosomes for each sample, indicating probable sex chromosome aneuploidy (see Methods). There are 652 samples with a probable sex chromosome aneuploidy (indicated by crosses). Locations of clusters of individuals with different putative karyotypes are indicated by Greek symbols: λ = X0 (or mosaic XX/X0), θ = XXX, α = XXY, and π = XYY. Counts of individuals in these regions are given in Supplementary Table 2. The colours indicate different combinations of self-reported sex, and sex inferred by Affymetrix (from the genetic data). For almost all samples (99.9%), the self-reported and the inferred sex are the same, but for a small number of samples (378) they do not match (see Supplementary Information for discussion).

for population structure in both marker and sample-based quality control (see Methods).

To identify poor quality markers, we used statistical tests designed primarily to check for consistency across experimental factors, such as array or batch (see Methods; Extended Data Table 4). As a result of these tests, we set to missing 0.97% of all the genotype calls made by Affymetrix. We identified poor quality samples using the metrics of missing rate and heterozygosity adjusted for population structure (Extended Data Fig. 1), as extreme values in one or both of these metrics can be indicators of poor sample quality due to, for example, DNA contamination[15]. We identified 968 such samples (0.2%), and provide this list to researchers.

Mismatches between self-reported sex of each individual, and sex inferred from the relative intensity of markers on the Y and X chromosomes[16], can be used as a way to detect possible sample mishandling or other types of clerical error. In a dataset of this size, some such mismatches would be expected due to transgender or intersex individuals, or instances of rare genetic variation, such as sex-chromosome aneuploidies[17]. Using information in the measured intensities of chromosomes X and Y (see Methods), we identified a set of 652 (0.134%) individuals with sex chromosome karyotypes that were putatively different from XY or XX (Fig. 2d, Supplementary Table 2).

The application of our quality control pipeline resulted in the released dataset of 488,377 samples and 805,426 markers from both arrays with the properties shown in Fig. 2a–c. A set of 588 pairs of experimental duplicates show very high genotype concordance, with mean 99.87% and minimum 99.39% of genotypes identical (Supplementary Fig. 13). We compared allele frequencies among UK Biobank participants with European ancestry to those estimated from an independent source, the Exome Aggregation Consortium (ExAC) database[18] at a set of 91,298 overlapping markers. We do not expect allele frequencies in the two studies to match exactly owing to subtle differences in the ancestral backgrounds of the individuals in each study, as well as differences in the sensitivity and specificity of the two technologies (exome sequencing and genotyping arrays). A small number of markers (around 300) have very different allele frequencies (see Supplementary Information section 2.4). This could be due to non-working probesets on the UK Biobank arrays or possibly annotation error on the UK Biobank arrays or in ExAC, or mapping errors in the sequence data in regions of more complex variation. Despite this, overall the allele frequencies are encouragingly similar ($r^2 = 0.93$) (Fig. 2c; Supplementary Fig. 4).

More than 110,000 rare markers (MAF < 0.01 in UK Biobank) were included on the two arrays used for the UK Biobank cohort[2]. Variants occurring at very low frequencies present a particular challenge for genotype calling using array technology. It can be challenging to distinguish a sample that genuinely has the minor allele, from one in which the intensities are in the tails of the distribution of those in the major homozygote cluster (Extended Data Fig. 2). A larger fraction of rare markers fail quality control tests compared to low frequency and common markers, but 84% still pass in all batches (Fig. 2b). We recommend researchers visually inspect cluster plots, similar to Supplementary Fig. 2, for markers of interest using a utility such as Evoker (https://github.com/wtsi-medical-genomics/evoker), especially for rare markers.

Ancestral diversity and cryptic relatedness

The genotype data provide a unique opportunity to study the diverse ancestral origins (Extended Data Table 3) of UK Biobank participants. Accounting for the ancestral background is essential both for epidemiological studies and genetic analyses, such as GWAS[19]. We used PCA to measure population structure within the UK Biobank cohort (see Methods). Figure 3a shows results for the first four principal components plotted in consecutive pairs (see also Extended Data Fig. 3 and Supplementary Figs. 6, 7). As expected, individuals with similar principal component scores have similar self-reported ethnic backgrounds. For example, the first two principal components separate out individuals with sub-Saharan African ancestry, European ancestry and east Asian ancestry. Individuals who self-report as mixed ethnicity tend to fall on a continuum between their constituent groups. Further principal components capture population structure at sub-continental geographic scales (Extended Data Fig. 3). Our PCA revealed population structure within the most common ethnic background category (88.26%), 'British' within the broader-level group 'white' (Supplementary Fig. 8). We used a combination of self-reported ethnic background and PCA results to provide researchers with a list of 409,728 individuals (84%) who have very similar ancestral backgrounds relative to the full cohort (see Methods).

Close relationships (for example, siblings) among UK Biobank participants were not recorded during the collection of other phenotypic information. This information can be important for epidemiological analyses[20], as well as in GWAS[21]. We used the genetic data to identify related individuals by estimating kinship coefficients for all pairs of samples, and report coefficients for pairs of relatives who we infer to be third-degree relatives or closer (see Methods). A total of 147,731 UK

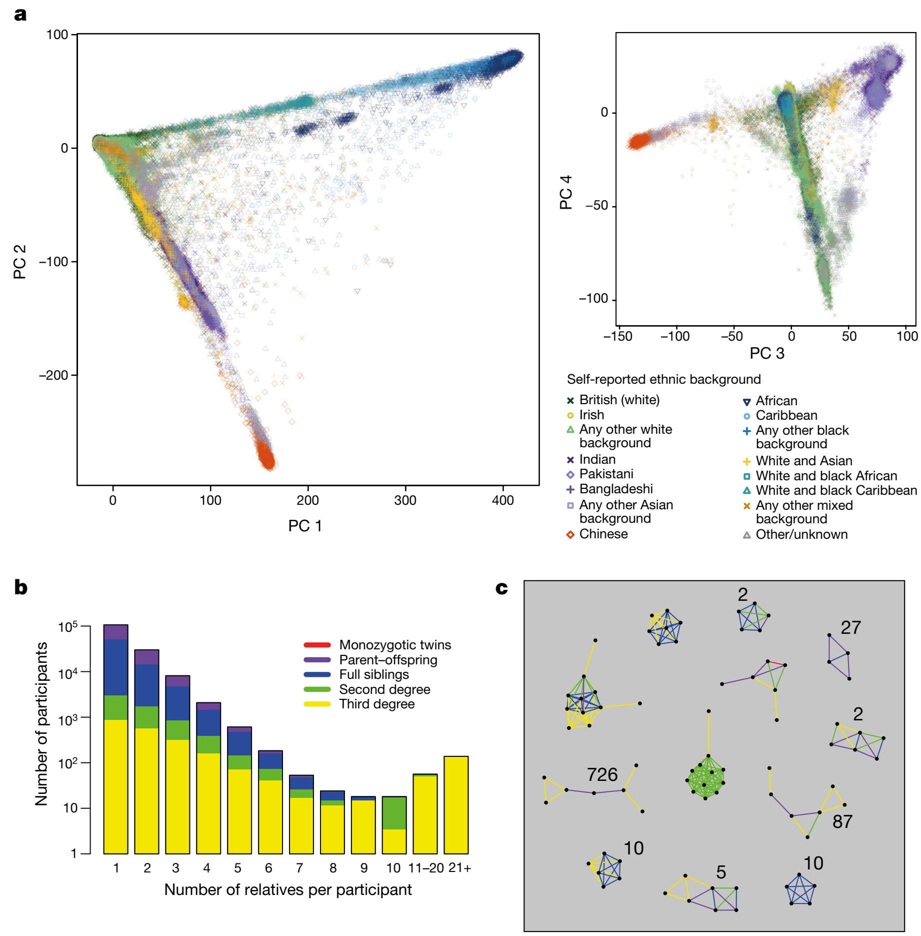


Fig. 3 | Ancestral diversity and familial relatedness. a, Each point represents a UK Biobank participant ($n = 488{,}377$ samples) and is placed according to their principal component (PC) scores in each of the top four principal components. Colours and shapes indicate the self-reported ethnic background of each individual. See Extended Data Table 3 for proportions in each category. **b**, Distribution of the number of relatives that participants have in the UK Biobank cohort. The height of each bar shows the count of participants ($\log_{10}$ scale) with the stated number of relatives. The colours indicate the proportions of each relatedness class within a bar. **c**, Examples of family groups within the UK Biobank cohort. Points represent participants, and coloured lines between points indicate their inferred relationship (for example, blue lines join full siblings). The integers show the total number of family networks in the cohort (if more than one) with that same configuration, ignoring third-degree pairs.

Biobank participants (30.3%) are inferred to be related (third degree or closer) to at least one other person in the cohort, and form a total of 107,162 related pairs (Extended Data Table 5). This is a surprisingly large number, and it is not driven solely by an excess of third-degree relatives. For example, the number of sibling pairs (22,666) is roughly twice as many as would theoretically be expected in a random sample (of this size) of the eligible UK population, after taking into account typical family sizes (Supplementary Table 4). The larger than expected number of related pairs could be explained by sampling bias due to, for example, an individual being more likely to agree to participate because a family member was also involved. Furthermore, if, as seems plausible, related individuals cluster geographically rather than being randomly located across the UK, the recruitment strategies of the UK Biobank assessment centres[22] will naturally tend to oversample related individuals.

Pairs of related individuals within the UK Biobank cohort form networks of related individuals. In most cases, these are of size two, but there are also many groups of size three or larger in the cohort (Fig. 3b), even when restricting to second-degree relatives or closer relative pairs. By considering the relationship types and the age and sex of the individuals within each family group, we identified 1,066 sets of trios (two parents and an offspring), which comprise 1,029 unique sets of parents and 37 quartets (two parents and two children).

There are 172 family groups with 5 or more individuals that are second-degree relatives or closer (Fig. 3c). One such group has 11 individuals who are all second-degree relatives of each other (half-siblings, grandparent/grandchild, or avuncular). Because all of the 55 pairs are second-degree relatives, at least 10 of them must be half-siblings with the same shared parent (see Supplementary Material). We confirmed that the shared parent must be their father because they do not all carry the same mitochondrial alleles, and the males all have the same Y chromosome alleles (data not shown).

Haplotype estimation and genotype imputation

We estimated haplotypes for the full cohort (pre-phasing), followed by haploid imputation[23]. For the pre-phasing step, we only used markers present on both the UK BiLEVE and UK Biobank Axiom arrays. We removed markers that failed quality control in more than one batch, had a greater than 5% overall missing rate, and had a MAF of less than 0.0001. We removed samples that were identified as outliers for heterozygosity and missing rate. These filters resulted in a dataset with 670,739 autosomal markers in 487,442 samples. Phasing on the autosomes was carried out using SHAPEIT3[24] (see Methods and https://jmarchini.org/software/). The 1000 Genomes phase 3 dataset[25] was used as a reference panel, predominantly to help with the phasing of samples with non-European ancestry. In a separate experiment that leveraged phase inferred from mother–father–child trios, we estimated a median phasing switch error rate of 0.229% (see Methods).

We used the Haplotype Reference Consortium (HRC)[26] data as the main imputation reference panel because it consisted of the largest

Table 1 | Association between HLA alleles and MS in UK Biobank and IMSGC cohort

HLA allele	Test	UK Biobank		IMSGC	
		OR (95% CI)	*P* value	OR (95% CI)	*P* value
*HLA-DRB1*15:01*	Additive effect	3.16 (2.81–3.54)	2.58×10^{-85}	3.92 (3.74–4.12)	$<1\times10^{-600}$
	Homozygote correction	0.67 (0.52–0.87)	2.32×10^{-3}	0.54 (0.47–0.61)	8.50×10^{-22}
*HLA-A*02:01*	Additive effect	0.69 (0.62–0.78)	2.30×10^{-10}	0.67 (0.64–0.70)	7.80×10^{-70}
	Homozygote correction	1.20 (0.89–1.62)	2.41×10^{-1}	1.26 (1.13–1.41)	3.30×10^{-05}
*HLA-DRB1*03:01*	Additive effect	1.21 (1.06–1.37)	3.39×10^{-3}	1.16 (1.10–1.22)	3.50×10^{-08}
	Homozygote correction	2.12 (1.53–2.94)	6.84×10^{-6}	2.58 (2.19–3.03)	1.30×10^{-30}
*HLA-DRB1*13:03*	Additive effect	2.10 (1.54–2.85)	2.36×10^{-6}	2.62 (2.32–2.96)	6.20×10^{-55}
*HLA-DRB1*08:01*	Additive effect	1.56 (1.21–2.01)	6.13×10^{-4}	1.55 (1.42–1.69)	1.00×10^{-23}
*HLA-B*44:02*	Additive effect	0.86 (0.74–0.98)	2.94×10^{-2}	0.78 (0.74–0.83)	4.70×10^{-17}
*HLA-B*38:01*	Additive effect	0.29 (0.13–0.65)	2.55×10^{-3}	0.48 (0.42–0.56)	8.00×10^{-23}
*HLA-B*55:01*	Additive effect	0.99 (0.75–1.31)	9.47×10^{-1}	0.63 (0.55–0.73)	6.90×10^{-11}
*HLA-DQA1*01:01*	Additive effect in the presence of *HLA-DRB1*15:01*	0.71 (0.56–0.90)	5.33×10^{-3}	0.65 (0.59–0.72)	1.30×10^{-17}
*HLA-DQB1*03:02*	Dominant effect	1.07 (0.92–1.25)	3.71×10^{-1}	1.30 (1.23–1.37)	1.80×10^{-22}
*HLA-DQB1*03:01*	Allelic interaction with *HLA-DQB1*03:02*	0.8 (0.53–1.20)	2.81×10^{-1}	0.60 (0.52–0.69)	7.10×10^{-12}

Evidence for association between HLA alleles and MS in UK Biobank compared to the IMSGC cohort. The UK Biobank association tests involved 1,501 self-reported cases and 409,724 controls; the IMSGC cohort involved 17,465 cases and 30,385 controls[31]. Thus, the UK Biobank analysis had significantly lower power than the IMSGC analysis, which is reflected in the reported *P* values and larger confidence interval (CI) estimates for the odds ratios (OR). Effect sizes for the UK Biobank were estimated jointly using the logistic regression model of the MHC reported by the IMSGC (with the exception of the two SNPs rs9277565 and rs2229029). As in the IMSGC analysis, the homozygote correction test indicates a departure from additivity. That is, if the odds ratio is <1 then the homozygous effect is smaller than under the additivity assumption and bigger if it is >1. Reported *P* values were calculated using the Wald test.

available set (64,976) of broadly European haplotypes at 39,235,157 SNPs. Supplementary Fig. 15 shows the results of a separate imputation experiment that shows that the HRC panel produces better imputation performance than the UK10K panel, especially at lower allele frequencies, and that the UK Biobank Axiom array performs favourably compared to other commercially available arrays.

We also imputed the UK Biobank using the merged UK10K and 1000 Genomes phase 3 reference panels[27], which has 87,696,888 bi-allelic markers. We combined this imputed data with that from the HRC panel, using the HRC imputation when a SNP was present in both panels. Imputation was carried out with the IMPUTE4 program (https://jmarchini.org/software/), which is a re-coded version of the haploid imputation functionality implemented in IMPUTE2[23] (see Methods). The result of the imputation process is a dataset with 93,095,623 autosomal SNPs, short indels and large structural variants in 487,442 individuals. We imputed an additional 3,963,705 markers on the X chromosome (Methods). The SNP database (dbSNP) reference SNP (rs) IDs were assigned to as many markers as possible using reference SNP ID lists available from the UCSC genome annotation database for the GRCh37 assembly of the human genome (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/).

Extended Data Fig. 4 shows the distribution of information scores on all markers in the imputed dataset. An information score of α in a sample of M individuals indicates that the amount of data at the imputed marker is approximately equivalent to a set of perfectly observed genotype data in a sample size of αM. The figure illustrates that most markers above 0.1% frequency have high information scores. Previous GWAS have tended to use a filter on information around 0.3 that roughly corresponds to an effective sample size of approximately 150,000. Thus, it may be possible to reduce the information score threshold and still obtain good power to detect associations.

We developed a new BGEN file format (v1.2; http://www.well.ox.ac.uk/~gav/bgen_format/bgen_format.html) and software library (BGEN; https://bitbucket.org/gavinband/bgen) to provide improved data compression, the ability to store phased haplotype data and random access to the data via use of a separate index file. Using this new format, the full imputed files require 2.1 Tb of file space. A new program (BGENIE; https://jmarchini.org/software) was built using the BGEN library to carry out fast multi-trait GWAS and phenome-wide association studies[28] (see Supplementary Information).

Imputation of classical HLA alleles

The major histocompatibility complex (MHC) on chromosome six is the most polymorphic region of the human genome and contains the largest number of genetic associations to common diseases[29]. We imputed HLA types at two-field (also known as four-digit) resolution for 11 classical HLA genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*) using the HLA*IMP:02 algorithm with a multi-population reference panel (Supplementary Tables 5 and 6)[30] and validated the accuracy using a cross-validation experiment. In a typical use, case accuracy was estimated at better than 96% across all loci (see Methods and Supplementary Tables 7, 8).

To demonstrate the utility of the HLA imputation, we performed association tests for diseases known to have HLA associations. We analysed 409,724 individuals in the white British ancestry subset (see Methods) and focused on 11 self-reported immune-mediated diseases with known HLA associations. For each disease in our analysis, we identified the HLA allele with the strongest evidence of association. In all cases these were consistent with previous reports (see Methods and Supplementary Table 9). We further replicated independent HLA associations in a single disease study of multiple sclerosis (MS) susceptibility by the International Multiple Sclerosis Genetics Consortium (IMSGC)[31]. Here we observed evidence of association and effect size estimates for HLA alleles that are concordant in direction and relative magnitude with those found in the IMSGC study, although in 11 out of 14 cases this was closer to 1, consistent with regression dilution bias arising from a low rate of phenotypic error (Table 1).

GWAS for standing height

To assess the potential of the directly genotyped and imputed data, we conducted a GWAS for standing height using 343,321 unrelated, European-ancestry UK Biobank participants (see Methods). We compared our results to a non-overlapping meta-analysis of 253,288 individuals of European ancestry carried out by the Genetic Investigation of Anthropometric Traits (GIANT) Consortium[32].

Reassuringly, the pattern of association signals is similar in both the UK Biobank and GIANT results (Fig. 4a–c), and the Z-scores at associated markers are highly correlated ($r^2 = 0.965$; Fig. 4e). The gain in power in the UK Biobank cohort is clear, with many loci reaching genome-wide significance ($P < 5\times10^{-8}$) in the UK Biobank but not in the GIANT study (Fig. 4d, Supplementary Fig. 16); and Z-scores for

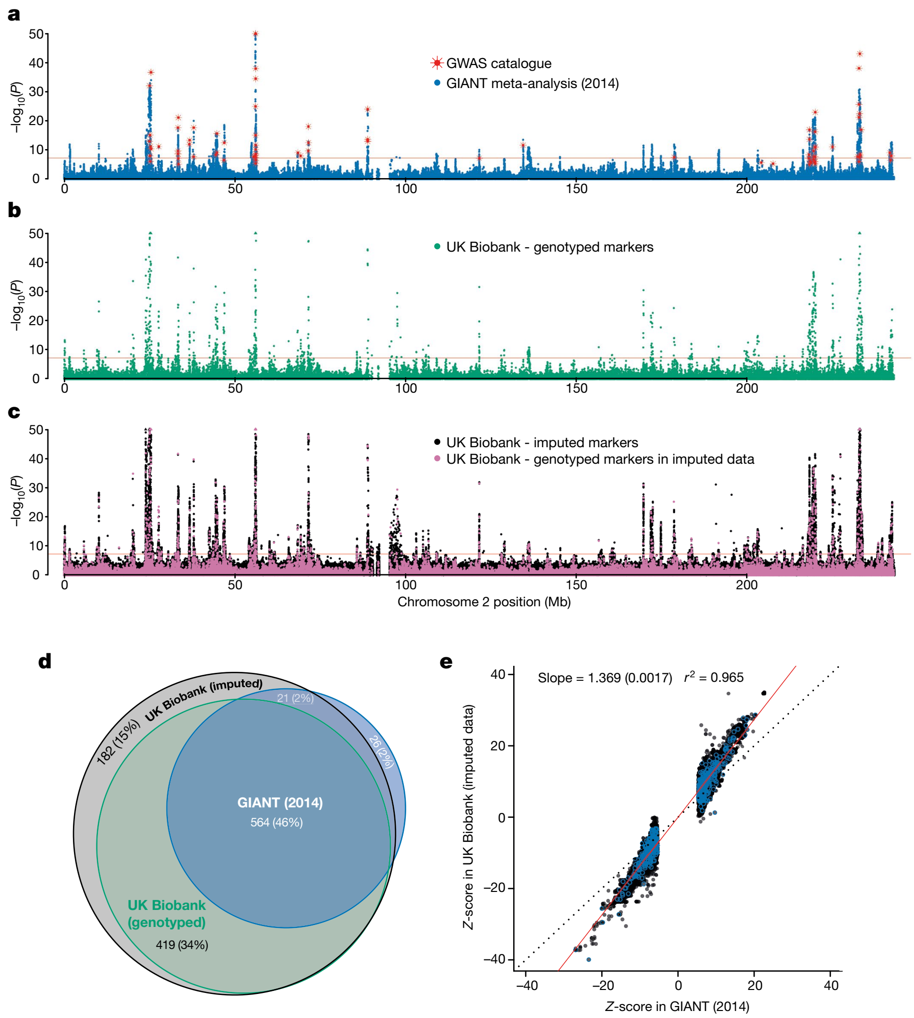


Fig. 4 | Association statistics for human height. Results (P values) of association tests between human height and genotypes using three different sets of data for chromosome 2. In **a**–**c**, P values are shown on the $-\log_{10}$ scale, capped at 50 for visual clarity and uncorrected for multiple comparisons. Markers with $-\log_{10}(P) > 50$ are plotted at 50 on the y axis and shown as triangles rather than dots. Horizontal red lines denote $P = 5 \times 10^{-8}$. **a**, Results for published meta-analysis by GIANT[32] ($n = 253{,}288$), with NCBI GWAS catalogue markers superimposed in red (plotted at the reported P values). **b**, Association statistics (from linear mixed model, see Methods) for UK Biobank markers in the genotype data ($n = 343{,}321$). **c**, Association statistics (from linear mixed model, see Methods) for UK Biobank markers in the imputed data ($n = 343{,}321$). Points coloured pink indicate genotyped markers that were used in pre-phasing and imputation. This means that most of the data at each of these markers comes from the genotyping assay. Black points (the vast majority, ~8 million) indicate fully imputed markers. **d**, Venn diagram of the results of counting the number of 1-Mb windows with at least one locus with $P < 5 \times 10^{-8}$ in the GIANT, UK Biobank genotyped and UK Biobank imputed datasets (see Methods). Percentages in brackets are the proportion of the union of such windows across all three data sources (1,215). There were only three windows contained in UK Biobank genotyped data and not the imputed data. **e**, Comparison of Z-scores in UK Biobank (y axis) and GIANT (x axis). Z-scores were calculated as effect size divided by standard error, but only for markers with $P < 5 \times 10^{-8}$ in GIANT, for a set of 575 associated regions, which we also used for the credible set analysis (see Methods). The marker with the smallest P value (in GIANT) within each region is highlighted with blue circles. The black dotted line shows $x = y$, and the red solid line shows the linear regression line estimated on these data. The standard error of the regression coefficient is shown in brackets. Pearson's correlation was used to calculate the r^2 value.

associated markers are systematically higher in UK Biobank (regression slope = 1.369, Fig. 4e). Regions of association in the UK Biobank show patterns of signal expected given the linkage disequilibrium structure and recombination rates in the region (see Extended Data Fig. 5 for an example).

To assess the effectiveness of UK Biobank genomic data for fine-mapping within associated loci, we computed 95% credible sets[33] for 575 regions that contain at least one genome-wide significant marker ($P < 5 \times 10^{-8}$) in both GIANT and the UK Biobank imputed data (see Methods). The number of markers we analysed in the UK Biobank (768,502) is considerably more than in GIANT (106,263), and this affects the resolution of any given associated region (Extended Data Fig. 6a). When considering all markers, the size of the credible set in UK Biobank is usually larger (median size = 8) than in GIANT (median size = 6), but the proportion of SNPs in the credible set of each region (Extended Data Fig. 6b) is generally smaller in UK Biobank (median proportion = 0.010) than in GIANT (median proportion = 0.047). By restricting to the markers in both studies (105,421) we find that the size of the 95% credible set is generally smaller in UK Biobank (median size = 4) than GIANT (median size = 6). The number of 95% credible sets that contain just 1 marker is 123 in UK Biobank and 76 in GIANT.

Conclusion

The interim release of the genetic data on approximately 150,000 participants in UK Biobank has already facilitated many papers exploring the links between human genetic variation and disease, and their connection with a wide range of environmental and lifestyle factors. The UK Biobank continues to grow with the addition of further phenotypic information and as researchers return the results of their analyses for UK

Biobank to share. Online resources are being developed for sharing the results of analyses using UK Biobank data, including the release of GWAS results for thousands of phenotypes (http://www.nealelab.is/uk-biobank) and the Oxford Brain Imaging Genetics server[28] (http://big.stats.ox.ac.uk/). We anticipate that the availability of the full genetic data for UK Biobank will result in a further step change in this productive research cycle. The UK Biobank is a powerful example of the immense value that can be achieved from large population scale studies that combine genetics with extensive and deep phenotyping and linkage to health records coupled with a strong data sharing policy. It is likely to herald a new era in which these and related resources drive and enhance understanding of human biology and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0579-z.

Received: 28 June 2018; Accepted: 6 September 2018;
Published online 10 October 2018.

1. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov*. **12**, 581–594 (2013).
2. The UK Biobank. *UK Biobank Axiom Array Content Summary* http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf (2014).
3. The UK Biobank. *Genotyping and Quality Control of UK Biobank, a Large-Scale, Extensively Phenotyped Prospective Resource* http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf (2015).
4. Young, A. I., Wauthier, F. & Donnelly, P. Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nat. Commun*. **7**, 12724 (2016).
5. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).
6. Wain, L. V. et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med*. **3**, 769–781 (2015).
7. Elliott, P. & Peakman, T. C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol*. **37**, 234–244 (2008).
8. Doherty, A. et al. Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank Study. *PLoS One* **12**, e0169649 (2017).
9. Miller, K. L. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci*. **19**, 1523–1536 (2016).
10. Petersen, S. E. et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank – rationale, challenges and approaches. *J. Cardiovasc. Magn. Reson*. **15**, 46 (2013).
11. Coffey, S. et al. Protocol and quality assurance for carotid imaging in 100,000 participants of UK Biobank: development and assessment. *Eur. J. Prev. Cardiol*. **24**, 1799–1806 (2017).
12. Harvey, N. C., Matthews, P., Collins, R., Cooper, C. & Group, U. B. M. A. Osteoporosis epidemiology in UK Biobank: a unique opportunity for international researchers. *Osteoporosis Int*. **24**, 2903–2905 (2013).
13. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. **12**, e1001779 (2015).
14. The UK Biobank. *Touchscreen Questionnaire Ordering, Validation and Dependencies* https://biobank.ctsu.ox.ac.uk/crystal/docs/TouchscreenQuestionsMainFinal.pdf (2018).
15. The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
16. Affymetrix. *Axiom Genotyping Solution Data Analysis Guide* http://tools.thermofisher.com/content/sfs/manuals/axiom_genotyping_solution_analysis_guide.pdf (2017).
17. Nielsen, J. & Wohlert, M. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Hum. Genet*. **87**, 81–83 (1991).
18. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
19. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet*. **36**, 512–517 (2004).
20. Shibata, K. et al. The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies. *Mol. Genet. Genomic Med*. **1**, 45–53 (2013).
21. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*. **1**, e32 (2005).
22. The UK Biobank. *UK Biobank: Protocol for a Large-Scale Prospective Epidemiological Resource* http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf (2007).
23. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet*. **44**, 955–959 (2012).
24. O'Connell, J. et al. Haplotype estimation for biobank-scale datasets. *Nat. Genet*. **48**, 817–820 (2016).
25. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
26. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet*. **48**, 1279–1283 (2016).
27. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun*. **6**, 8111 (2015).
28. Elliott, L. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat. Commun*. **9**, 1470 (2018).
29. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. **42**, D1001–D1006 (2014).
30. Dilthey, A. et al. Multi-population classical HLA type imputation. *PLOS Comput. Biol*. **9**, e1002877 (2013).
31. The International Multiple Sclerosis Genetics Consortium. Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat. Genet*. **47**, 1107–1113 (2015).
32. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet*. **46**, 1173–1186 (2014).
33. The Wellcome Trust Case Control Consortium et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet*. **44**, 1294–1301 (2012).

Acknowledgements We acknowledge Wellcome Trust Core Awards 090532/Z/09/Z and 203141/Z/16/Z and grants 095552/Z/11/Z (to P.D.), 100956/Z/13/Z (to G.M.) and 100308/Z/12/Z (to A.C.). J.M. is supported by European Research Council grant 617306. S.L. is supported by Australian NHMRC Career Development Fellowship 1053756. The sample processing and genotyping was supported by the National Institute for Health Research, Medical Research Council, and British Heart Foundation. We thank the Research Computing Core at the Wellcome Centre for Human Genetics for assistance with the computational workload. We thank Affymetrix for discussions concerning quality control. We thank A. Young, A. Dilthey and L. Moutsianas for their assistance with aspects of the data analysis. We acknowledge UK Biobank co-ordinating centre staff for their role in extracting the DNA for this project. We thank M. Kuzma-Kuzniarska (http://mybioscience.org/) for Fig. 1.

Reviewer information *Nature* thanks E. Banks, M. Boehnke, B. Pasaniuc, D. MacArthur and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.Y., N.A. and M.E. coordinated data availability. Data analysis was performed as follows: quality control analysis (C.B., C.F., D.P. and S.W.), HLA imputation and association testing (A.C., A.M. and D.V), phasing, imputation, file formats and multiple trait analysis (O.D., J.O., G.B., K.S., L.T.E. and J.M.) and GWAS testing (C.B., C.F. and J.M.). Supervision of these activities was provided by G.M., P.D. and J.M. C.B., C.F., A.C., S.L., N.A., G.M., P.D. and J.M. wrote the paper.

Competing interests J.M. is a founder and director of Gensci Ltd. P.D., G.M. and S.L. are partners in Peptide Groove LLP. G.M. and P.D. are founders and directors of Genomics Plc. The remaining authors declare no competing financial interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0579-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0579-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.M.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data collection, sample retrieval, DNA extraction and genotype calling. Ethics approval for the UK Biobank study was obtained from the North West Centre for Research Ethics Committee (11/NW/0382). Blood samples were collected from participants on their visit to a UK Biobank assessment centre and the samples are stored at the UK Biobank facility in Stockport, UK[7]. Over a period of 18 months samples were retrieved, DNA was extracted, and 96-well plates of 94×50-μl aliquots were shipped to Affymetrix Research Services Laboratory for genotyping. Special attention was paid in the automated sample retrieval process at UK Biobank to ensure that experimental units such as plates or timing of extraction did not correlate systematically with baseline phenotypes such as age, sex, and ethnic background, or the time and location of sample collection. Full details of the UK Biobank sample retrieval and DNA extraction process were described previously[34].

On receipt of DNA samples, Affymetrix processed samples on the GeneTitan Multi-Channel (MC) Instrument in 96-well plates containing 94 UK Biobank samples and two control samples from the 1000 Genomes Project[25]. Genotypes were then called from the array intensity data, in units called 'batches' which consist of multiple plates. Across the entire cohort, there were 106 batches of 4,700 UK Biobank samples each (Supplementary Information, Supplementary Table 12). Following the earlier interim data release, Affymetrix developed a custom genotype calling pipeline that is optimized for biobank-scale genotyping experiments, which takes advantage of the multiple-batch design[35]. This pipeline was applied to all samples, including the 150,000 samples that were part of the interim data release. Consequently, some of the genotype calls for these samples may differ between the interim data release and this final data release (see below).

Routine quality checks were carried out during the process of sample retrieval, DNA extraction[36], and genotype calling[37]. Any sample that did not pass these checks was excluded from the resulting genotype calls. The custom-designed arrays contain a number of markers that had not been previously typed using Affymetrix genotype array technology. As such, Affymetrix also applied a series of checks to determine whether the genotyping assay for a given marker was successful, either within a single batch, or across all samples. Where these newly attempted assays were not successful, Affymetrix excluded the markers from the data delivery (see Supplementary Information for details).

Marker-based quality control. We identified poor quality markers using statistical tests designed primarily to check for consistency of genotype calling across experimental factors. Specifically we tested for batch effects, plate effects, departures from Hardy–Weinberg equilibrium, sex effects, array effects, and discordance across control replicates. See Supplementary Information for the details of each test, and Supplementary Fig. 3 for examples of affected markers. For markers that failed at least one test in a given batch, we set the genotype calls in that batch to missing. We also provide a flag in the data release that indicates whether the calls for a marker have been set to missing in a given batch. If there was evidence that a marker was not reliable across all batches, we excluded the marker from the data altogether. To attenuate population structure effects, we applied all marker-based quality control tests using a subset of 463,844 individuals with estimated European ancestry. We identified these individuals from the genotype data before conducting any quality control by projecting all the UK Biobank samples on to the two major principal components of four 1000 Genomes populations (CEU, YRI, CHB and JPT)[25]. We then selected samples with principal component scores falling in the neighbourhood of the CEU cluster (Supplementary Information).

Sample-based quality control. We identified poor quality samples using the metrics of missing rate and heterozygosity computed using a set of 605,876 high quality autosomal markers that were typed on both arrays (see Supplementary Information for criteria). Extreme values in one or both of these metrics can be indicators of poor sample quality due to, for example, DNA contamination[15]. The heterozygosity of a sample—the fraction of non-missing markers that are called heterozygous—can also be sensitive to natural phenomena, including population structure, recent admixture and parental consanguinity. We took extra measures to avoid misclassifying good quality samples because of these effects. For example, we adjusted heterozygosity for population structure by fitting a linear regression model with the first six principal components in a PCA as predictors (Extended Data Fig. 1). Using this adjustment we identified 968 samples with unusually high heterozygosity or >5% missing rate (Supplementary Information). A list of these samples is provided as part of the data release.

We also conducted quality control specific to the sex chromosomes using a set of 15,766 high quality markers on the X and Y chromosomes. Affymetrix infers the sex of each individual based on the relative intensity of markers on the Y and X chromosomes[16]. Sex is also reported by participants, and mismatches between these sources can be used as a way to detect sample mishandling or other kinds of clerical error. However, in a dataset of this size, some such mismatches would be expected due to transgender individuals, or instances of real (but rare) genetic variation, such as sex-chromosome aneuploidies[17]. Affymetrix genotype calling on the X and Y chromosomes allows only haploid or diploid genotype calls, depending on the inferred sex[16]. Therefore, cases of full or mosaic sex chromosome aneuploidies may result in compromised genotype calls on all, or parts of, the sex chromosomes (but not affect the autosomes). For example, individuals with karyotype XXY will probably have poorer quality genotype calls on the pseudo-autosomal region (PAR) of the X chromosome, as they are effectively triploid in this region. Using information in the measured intensities of chromosomes X and Y, we identified a set of 652 (0.134%) individuals with sex chromosome karyotypes putatively different from XY or XX (Fig. 2d, Supplementary Table 2). The list of samples is provided as part of the data release. Researchers wanting to identify sex mismatches should compare the self-reported sex and inferred sex data fields.

We did not remove samples from the data as a result of any of the above analyses, but rather provide the information as part of the data release. However, we excluded a small number of samples (835 in total) that we identified as sample duplicates (as opposed to identical twins, see Supplementary Information) or were probably involved in sample mishandling in the laboratory (~10), as well as participants who asked to be withdrawn from the project before the data release.

Comparison of interim and final release data. Subsequent to the interim release of genotypes (May 2015) for approximately 150,000 UK Biobank participants improvements were made to the genotype calling algorithm[35] and quality control procedures. We therefore expect to observe some changes in the genotype calls and missing data profile of samples included in both the interim data release and this final data release. Discordance among non-missing markers is very low (mean 6.7×10^{-5}; Supplementary Fig. 1); and for each sample there are 24,500 genotype calls (on average) that were missing in the interim data, but which have non-missing calls in this release. This is much smaller in the reverse direction, with 500 calls, on average, missing in this release but not missing in the interim data, so there is an average net gain of 24,000 genotype calls per sample.

Principal component analysis. We computed principal components using an algorithm (fastPCA[38]) that performs well on datasets with hundreds of thousands of samples by approximating only the top n principal components that explain the most variation, in which n is specified in advance. We computed the top 40 principal components using a set of 407,219 unrelated, high quality samples and 147,604 high quality markers pruned to minimise linkage disequilibrium[39]. We then computed the corresponding principal component-loadings and projected all samples onto the principal components, thus forming a set of principal component scores for all samples in the cohort (Supplementary Information).

White British ancestry subset. Researchers may want to only analyse a set of individuals with relatively homogeneous ancestry to reduce the risk of confounding due to differences in ancestral background. Although the UK Biobank cohort includes a large number of participants from a wide range of ethnic backgrounds, such analysis is feasible without compromising too much in sample size because most participants in the UK Biobank cohort report their ethnic background as 'British', within the broader-level group 'white' (88.26%). Our PCA revealed population structure even within this category (Supplementary Fig. 8), so we used a combination of self-reported ethnic background and genetic information to identify a subset of 409,728 individuals (84%) who self-report as 'British' and who have very similar ancestral backgrounds based on results of the PCA (Supplementary Information). Fine-scale population structure is known to exist within the UK but methods for detecting such subtle structure[40] available at the time of analysis are not feasible to apply at the scale of the UK Biobank. The white British ancestry subset may therefore still contain subtle structure present at sub-national scales.

Kinship coefficient estimation. We used an estimator implemented in the software, KING[41], as it is robust to population structure (that is, does not rely on accurate estimates of population allele frequencies) and it is implemented in an algorithm efficient enough to consider all pairs ($\sim 1.2 \times 10^{11}$) in a practicable amount of time. As noted by the authors of KING, we found that recent admixture (for example, 'mixed' ancestral backgrounds) tended to inflate the estimate of the kinship coefficient, as the estimator assumes Hardy–Weinberg equilibrium among markers with the same underlying allele frequencies within an individual. We alleviated this effect by only using a subset of markers that are only weakly informative of ancestral background (Supplementary Information, Supplementary Fig. 12). We also excluded a small fraction of individuals (977) from the kinship estimation, as they had properties (for example, high missing rates) that would lead to unreliable kinship estimates (Supplementary Information). We called relationship classes for each related pair using the kinship coefficient and fraction of markers for which they share no alleles (IBS0). See Supplementary Information section S3.7 for details.

To ensure we were not overestimating the number of related pairs, we inferred related pairs (within a subset of the data) using a different inference method implemented in PLINK ('-genome' command; https://www.cog-genomics.org/plink2) and confirmed 100% of the twins, parent-offspring and sibling pairs, and 99.9% of pairs overall (Supplementary Information).

Haplotype estimation. Haplotype estimation (phasing) was carried out using SHAPEIT3 in chunks of 15,000 markers, with an overlap of 250 markers between chunks. Each chunk used 4 cores per job and $S = 200$ copying states. Chunks were ligated using a modified version of the hapfuse program (https://bitbucket.org/wkretzsch/hapfuse/src).

We assessed the accuracy of the phasing in a separate experiment by taking advantage of mother-father-child trios that were identified in the UK Biobank cohort. This family information can be used to infer the phase of a large number of markers in the trio parents. These family-inferred haplotypes were used as a truth set, as is common in the phasing literature. The parents of each trio were removed from the dataset and then haplotypes were estimated across chromosome 20 in a single run of SHAPEIT3. This dataset consisted of 16,175 autosomal markers. The inferred haplotypes were then compared to the truth set using the switch error metric. Using a set of 696 trios with self-reported ethnic background 'British' (within the broader-level group 'white') and no other twins or first- or second-degree relatives in the UK Biobank dataset, we estimated a median switch error rate of 0.229%. We also used a subset of 397 of these trios that also had no third-degree relatives and obtained a median switch error rate of 0.234%. These error rates are similar to those produced by other phasing methods that can handle data at this scale[42,43]. Investigations on the effect of sample size on phasing performance and downstream imputation performance suggest that differences between methods will have negligible effect on genotype imputation and GWAS[42].

Imputation. To facilitate fast imputation of all 500,000 samples, we re-coded IMPUTE2[23] to focus exclusively on the haploid imputation needed when samples have been pre-phased. This new version of the program is referred to as IMPUTE4 (see https://jmarchini.org/software/), but uses exactly the same hidden Markov model within IMPUTE2, and produces identical results to IMPUTE2 when run using all reference haplotypes as hidden states (data not shown). To reduce RAM usage and increase speed we use compact data structures that store the indices of haplotypes carrying the non-reference allele at variant sites in the reference panel. Not only is this data structure compact, but at each stage of the forward-backward algorithm it also allows the calculations involving the emission part of the hidden Markov model to sum only over just the subset of haplotypes that carrying the non-reference allele in an efficient way. A further increase in speed is obtained by only calculating the marginal copying probabilities at those sites common to the target and reference datasets, and then linearly interpolating these for SNPs in-between those sites that need to be imputed. Imputation was carried out in chunks of approximately 50,000 imputed markers with a 250 kb buffer region and on 5,000 samples per compute job. The combined processing time per sample for the whole genome was approximately 10 min.

Haplotype estimation and genotype imputation on the X chromosome. For haplotype estimation on the X chromosome genotype data we applied the same filtering steps as the autosomal genotype data, with some additional filters. For both the sex-specific region and the pseudo-autosomal regions (PAR), samples were excluded which were identified as having a likely sex chromosome aneuploidy (see above). For the PAR, we additionally excluded samples with a missing rate of >5% among markers in the PAR. For the sex-specific region of chromosome X, this resulted in a dataset of 16,601 markers and 486,790 samples. For the PAR this resulted in a dataset of 1,239 markers and 486,476 samples. Haplotype estimation and genotype imputation was carried out on the two pseudo-autosomal regions and the non-pseudo autosomal region separately, and using the same methods and reference datasets used for the autosomes.

HLA imputation and validation. For each individual we defined the HLA genotype at each locus as the pair of alleles with maximum posterior probability as reported by HLA*IMP:02. We performed association analysis (see, for example, ref. [31]) for HLA alleles and each disease using logistic regression. The risk model (additive, dominant, recessive or general), as described previously[31], was used to enable comparison of effect size estimates. For validation and further details, see Supplementary Information section S5. We repeated the analysis, setting genotypes with a maximum posterior probability of <0.7 to missing. No significant differences were observed compared to the full analysis (data not shown). As a negative control, we ran association analyses in the HLA region with imputed HLA alleles for type 2 diabetes (2,849 cases) and myocardial infarction (9,725 cases) in a total of 409,724 individuals and we found no significant associations (all $P > 2.40 \times 10^{-4}$, the Bonferroni corrected level of association) with any HLA alleles, which is consistent with the lack of associations in the HLA region in recent analyses of each phenotype[44,45]

We estimated the accuracy of the imputation process using fivefold cross-validation in the reference panel samples. For samples of European ancestry, the estimated four-digit accuracy for the maximum posterior probability genotype is above 93.9% for all 11 loci (Supplementary Table 7). This accuracy improved to above 96.1% for all 11 loci after restricting to HLA allelic variant calls with a posterior probability greater than 0.70. This resulted in call rates above 95.1% for all loci (Supplementary Table 8).

GWAS for standing height. We conducted the GWAS for standing height using the directly genotyped and imputed data in the form that they are made available to researchers, but with a subset of samples. Specifically, we only included samples with all of the following properties: (i) imputation was carried out on them; (ii) in the white British ancestry subset (see above); and (iii) the inferred sex matches the self-reported sex. From this group we selected a set of 344,397 unrelated individuals (Supplementary Information). For standing height, a further 1,076 individuals were excluded owing to missing values for the phenotype, leaving a total of 343,321 for association testing.

We used the software BOLT-LMM (v2.2)[46] to look for evidence of statistical association between each marker and standing height. We report association statistics based on a linear mixed model (BOLT-LMM-inf), with the following covariates: (i) array (UK BiLEVE Axiom Array or UK Biobank Axiom Array); (ii) sex (inferred); (iii) age when attended UK Biobank assessment centre; and (iv) principal components 1–20.

The principal components scores were computed using only individuals within the white British ancestry subset, but otherwise with the same method as described above. We conducted tests using the genotype and imputed data files separately.

Example of association region in standing height GWAS. Extended Data Fig. 5 shows an example of an associated region on chromosome 2. Correlations (r^2) between markers in this region show a pattern that is as expected in the context of linkage disequilibrium, and the local recombination rates. The stripe-like pattern of the association statistics is indicative of multiple mutations occurring on similar branches of the genealogical tree underlying the data, which are probably linked to varying degrees with the causal marker(s). The correlation between the most associated marker and all other markers in the region drops off sharply around the small peak in recombination[47] to the right of the most significantly associated marker. Notably, this marker was imputed from the genotypes, which points to the success of the imputation in this study, and in general, to the value of imputing millions more markers. Human height is a highly polygenic trait, so provided an opportunity to examine many such regions of association, and other regions that we visually examined showed similar patterns.

Comparison of GIANT and UK Biobank GWAS results. For Fig. 4d, e and the credible set analysis we used autosomal markers only, and filtered markers in each data source such that MAF > 0.001 (defined in the GWAS population), and Info score > 0.3 in the UK Biobank imputed data. There were 16,443,622 such markers in UK Biobank imputed data, 703,946 in the UK Biobank genotyped data, and 2,546,872 in GIANT.

For a given phenotype, the 95% credible set in a region of association is the smallest set of markers that together have 95% posterior probability of containing the marker causally associated with the phenotype. We found credible sets for standing height using the method described previously[33] and summarize the results in Extended Data Fig. 6. It is important to note that this approach is based on a model in which there is exactly one causal marker in the region and genotypes for that marker are available in the data. Our results should therefore be considered as indicative of a more detailed analysis where, for example, the regions are first analysed to distinguish independent association signals.

In our analysis, we first defined a set of 575 non-overlapping regions associated with standing height using a procedure based on that used previously[15] (see Supplementary Information). For each study, we carried out two separate analyses to find credible sets in these regions: (A) using all the markers in each study (768,502 in UK Biobank imputed data; 106,263 in GIANT); and (B) using only those markers in both studies (105,421).

For each marker in each study, we computed a Bayes factor in favour of association with standing height using the effect sizes and standard errors, and 0.2^2 as the prior[33] on the variance of the effect sizes. To ensure the effect sizes were on the same scale in both studies we scaled UK Biobank effect sizes and standard errors by the standard deviation of the residuals of the measured phenotype (standing height) after regressing out the covariates used in the GWAS. We then confirmed that the effect size estimates for overlapping markers were comparable between the two studies.

If there is exactly one causal marker in the region and genotypes for that marker are available in the data, then the posterior probability that a marker i drives the association signal in the region r is given by:

$$\pi_{ir} = \frac{\mathrm{BF}_{ir}}{\Sigma_k \mathrm{BF}_{kr}}$$

where BF_{kr} is the Bayes factor for marker i in the r region[33]. The 95% credible set for a region is found by going down the list of markers ordered from highest to lowest posterior probability and stopping when the cumulative posterior reaches 0.95.

We assessed the sensitivity of our results to the choice of prior by conducting the same analyses using a much smaller prior (0.02^2) and much larger prior (20^2). We found that overall the choice of prior had little effect on the results. Specifically for values we report in the main text, the median credible set sizes were unaffected in all analyses. For the larger prior, the number of single-marker credible sets was unaffected except for analysis B in UK Biobank (from 123 to 122), and the median proportion of markers in the credible set was unaffected in all analyses. For the smaller prior, the number of single-marker credible sets only changed for analysis A, going from 78 to 75 in GIANT, and 85 to 86 in UK Biobank, and the median proportion of markers in the credible set increased slightly in all analyses (maximum increase from 0.047 to 0.051).

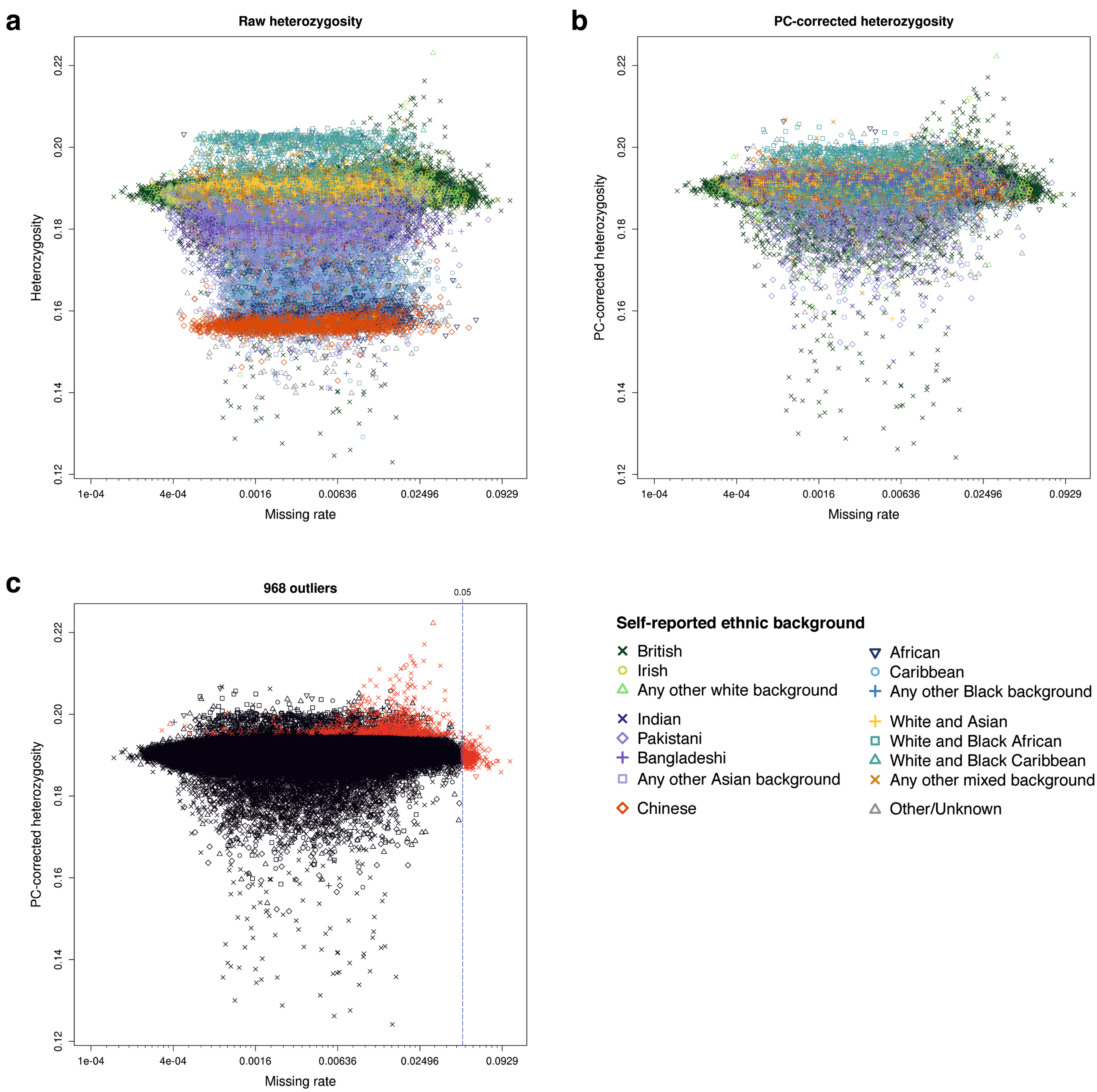
Code availability. Genotype imputation was carried out using IMPUTE4.0. Pre-compiled binaries for the latest version of IMPUTE4 are available at https://jmarchini.org/software/. This software is licensed free for use by researchers at academic institutions. The BGEN library source code is available at https://bitbucket.org/gavinband/bgen. BGENIE is built using this library. Pre-compiled binaries for the latest version of BGENIE are available at https://jmarchini.org/software/. This software is currently licensed free for use by researchers at academic institutions. Commercial organizations wishing to use IMPUTE4 or BGENIE must enquire about a licence from the University of Oxford.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

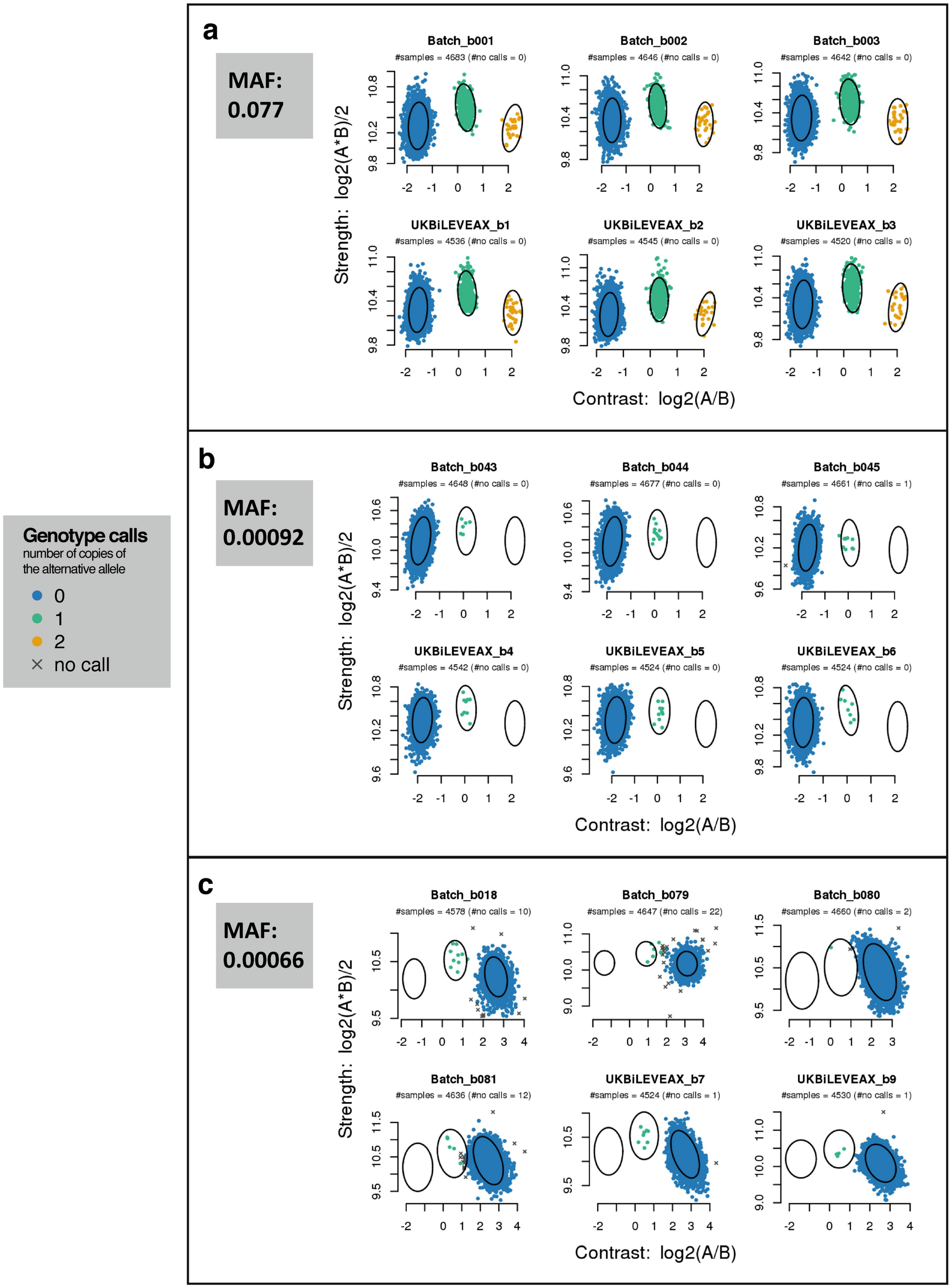
Data availability

The genetic and phenotype datasets generated by UK Biobank analysed during the current study are available via the UK Biobank data access process (see http://www.ukbiobank.ac.uk/register-apply/). Detailed information about the genetic data available from UK Biobank is available at http://www.ukbiobank.ac.uk/scientists-3/genetic-data/ and http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100314. The exact number of samples with genetic data currently available in UK Biobank may differ slightly from those described in this paper.

34. Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* **18**, 26 (2017).
35. Affymetrix. *UKB_WCSGAX: UK Biobank 500K Samples Genotyping Data Generation by the Affymetrix Research Services Laboratory* http://biobank.ndph.ox.ac.uk/showcase/docs/affy_data_generation2017.pdf (2017).
36. UK Biobank. *Genotyping of 500,000 UK Biobank Participants: Description of Sample Processing Workflow and Preparation of DNA for Genotyping* https://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_sample_workflow.pdf (2015).
37. Affymetrix. *UKB_WCSGAX: UK Biobank 500K Samples Processing by the Affymetrix Research Services Laboratory* http://biobank.ndph.ox.ac.uk/showcase/docs/affy_lab_process2017.pdf (2017).
38. Galinsky, K. J. et al. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
39. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 132–135, author reply 135–139 (2008).
40. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
41. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
42. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
43. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
44. Webb, T. R. et al. Systematic evaluation of pleiotropy identifies 6 further loci associated with coronary artery disease. *J. Am. Coll. Cardiol.* **69**, 823–836 (2017).
45. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
46. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
47. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
48. Galante, J. et al. The acceptability of repeat Internet-based hybrid diet assessment of previous 24-h dietary intake: administration of the Oxford WebQ in UK Biobank. *Br. J. Nutr.* **115**, 681–686 (2016).

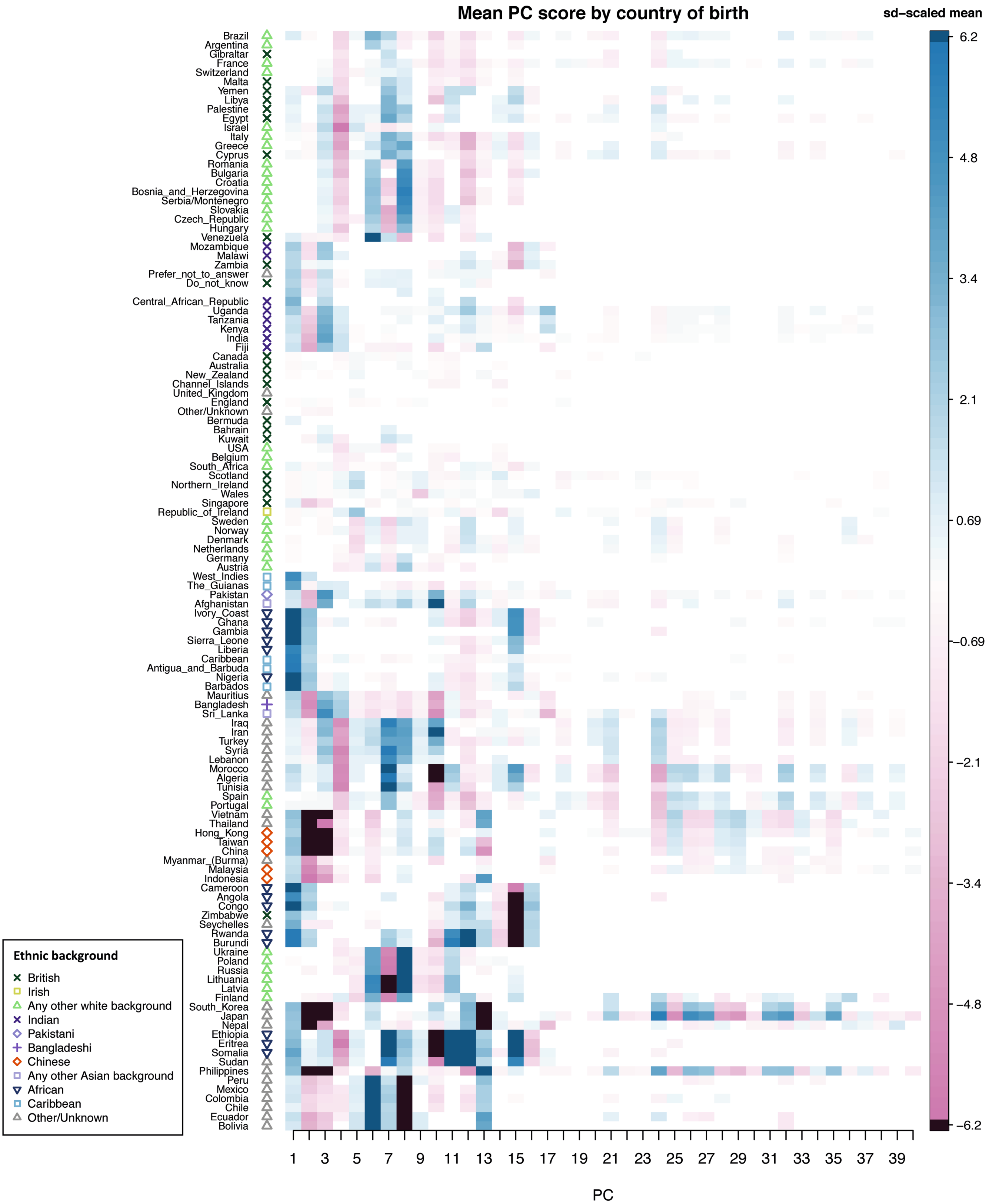


Extended Data Fig. 1 | Summary of sample-based quality control. **a**–**c**, The three plots show heterozygosity and missing rates, which we used to flag poor quality samples ($n = 488{,}377$ samples). Panels **a** and **b** show heterozygosity for each sample before and after, respectively, correcting for ancestral background using principal components. The symbols (shapes and colours) indicate the self-reported ethnic background of each participant. Panel **c** shows the set of 968 samples we flagged as outliers (in red), and all other samples (in black), with shapes the same as for the other two plots. The vertical line shows the threshold we used to call samples as outliers on missing rate. In all plots missing rate data are transformed to the logit scale, but with the axis annotated with the original values.

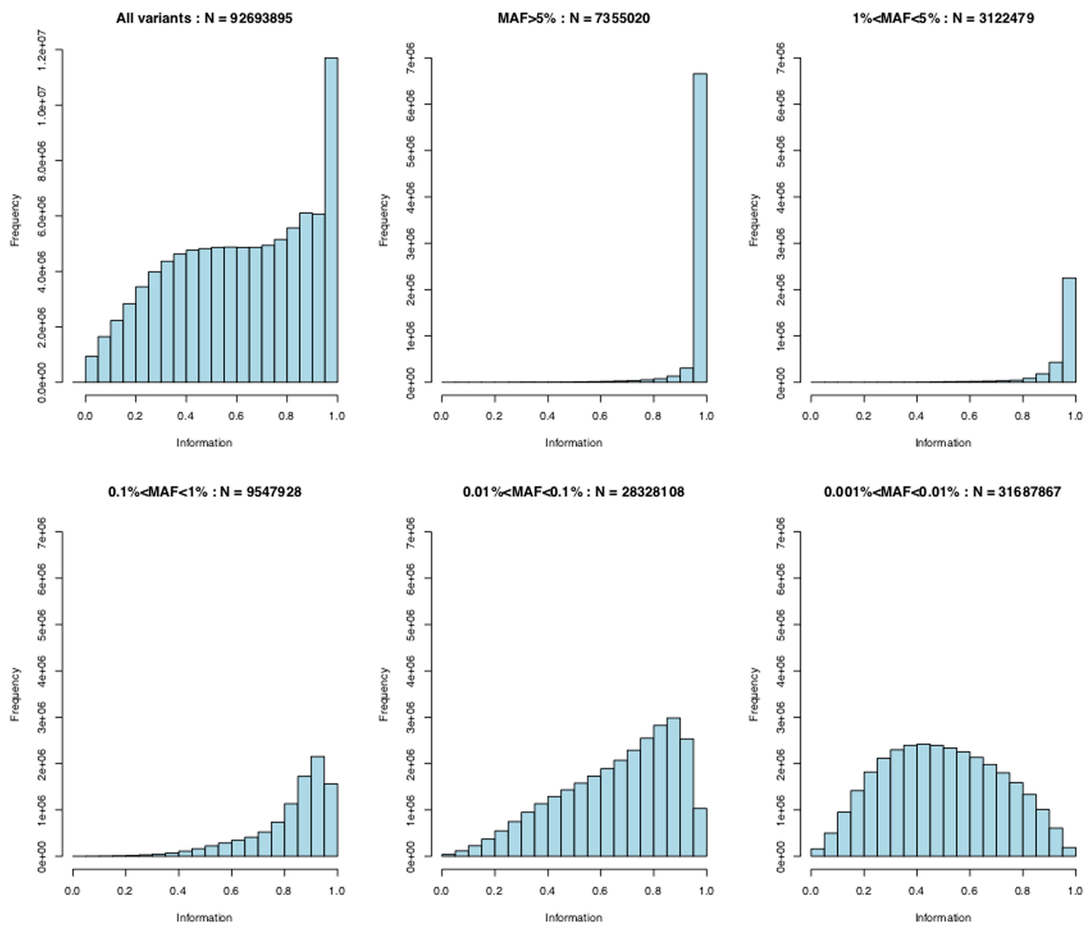


Extended Data Fig. 2 | See next page for caption.

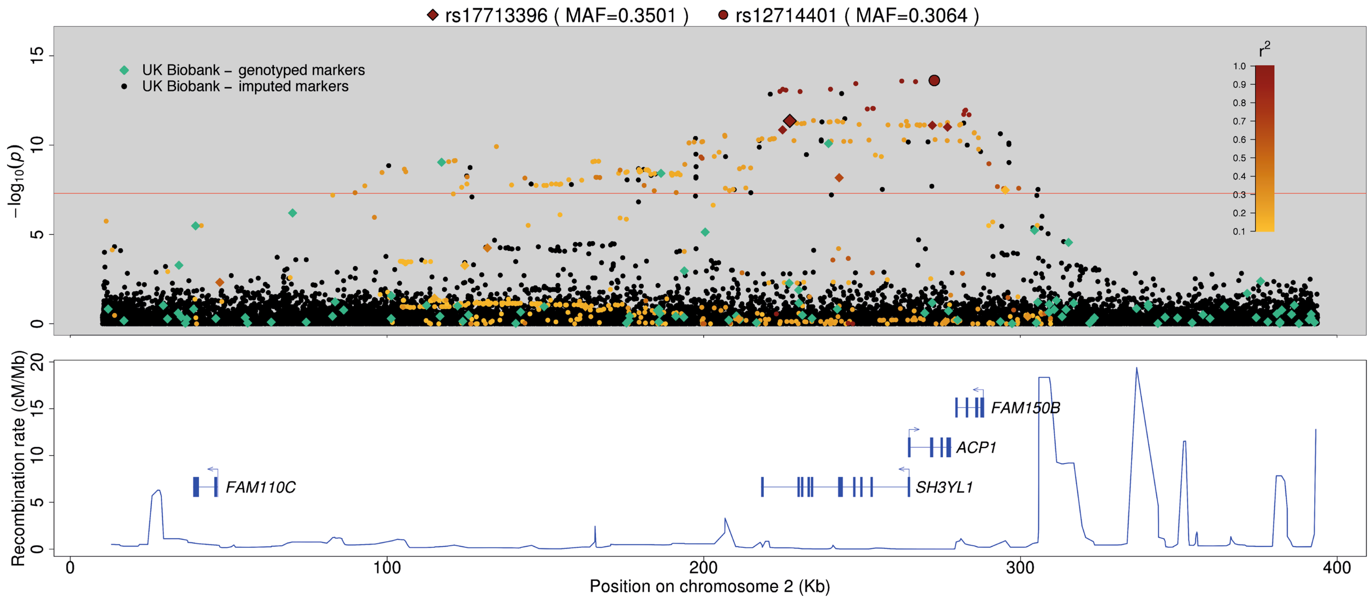
Extended Data Fig. 2 | Examples of intensity data and genotype calls for markers of different allele frequencies. Each sub-figure shows intensity data for a single marker within six different batches. Batches labelled with the prefix 'UKBiLEVEAX' contain only samples typed using the UK BiLEVE Axiom array, and those with the prefix 'batch' contain only samples typed using the UK Biobank Axiom array. Each point represents one sample and is coloured according to the inferred genotype at the marker. The *x* and *y* axes are transformations of the intensities for probe sets targeting each of the alleles 'A' and 'B' (see Supplementary Information for definition of probe set). The ellipses indicate the location and shape of the posterior probability distribution (two-dimensional multivariate normal) of the transformed intensities for the three genotypes in the stated batch. That is, each ellipse is drawn such that it contains 85% of the probability density. See Affymetrix *Axiom Genotyping Solution Data Analysis Guide*[16] for more details of Affymetrix genotype calling. The MAF of each of the markers is computed using all samples in the released UK Biobank genotype data. **a**, A marker with a MAF of 0.077 with well-separated genotype clusters. **b**, Intensities for a marker with a MAF of 0.00092 with well-separated genotype clusters. As would be expected under Hardy–Weinberg equilibrium, there are no instances of samples with the minor homozygote genotype. **c**, Intensities for a marker with a MAF of 0.00066, and in which the heterozygote cluster is not well separated from the large major homozygote cluster in some batches, making it more difficult to call the heterozygous genotypes confidently.



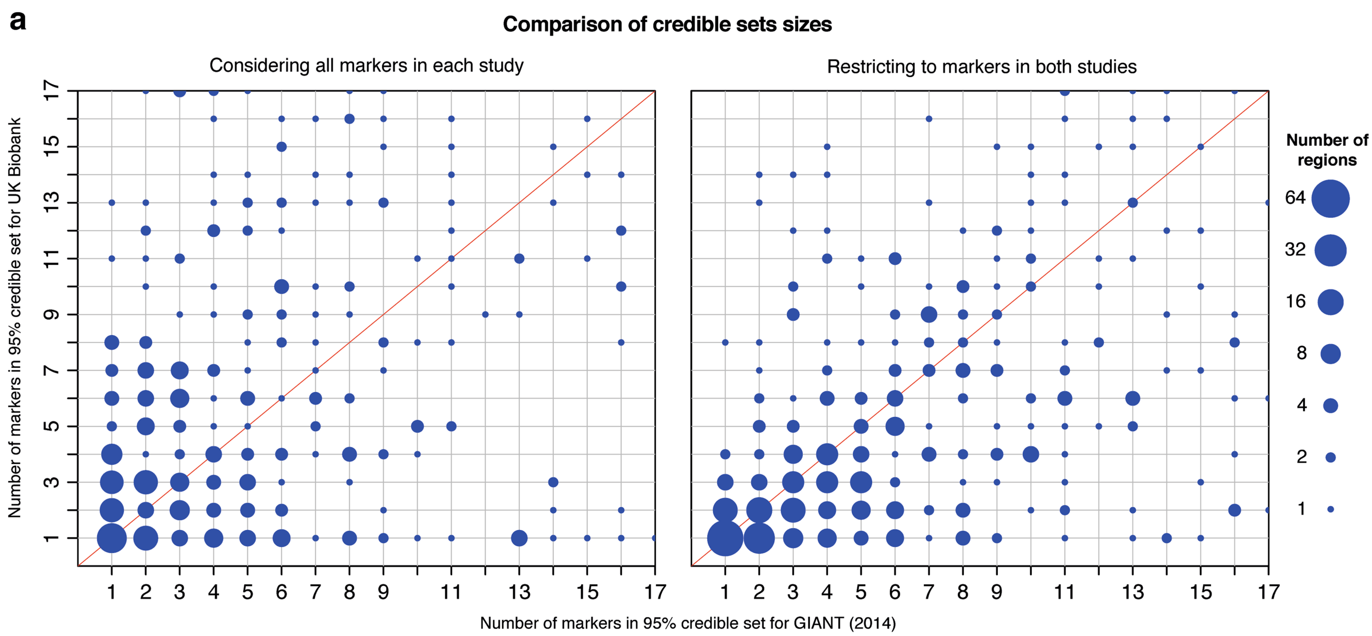
Extended Data Fig. 3 | Mean principal component scores for each self-reported country of birth. Each column shows one principal component and each element is the mean principal component score for individuals born in the labelled country, scaled by the standard deviation of the scores for that principal component. Elements in each column are only coloured if the country has a non-zero coefficient ($P < 10^{-5}$; two-sided t-test) in a linear model with country of birth as predictor and principal component scores as outcome ($n = 487{,}848$ samples). Countries (rows) have been ordered using hierarchical clustering ('hclust' function in R). The symbols next to each country label indicate the most common ethnic background category among the participants born in that country. For example, the most common self-reported ethnic background of participants born in Sri Lanka is 'Any other Asian background'. Countries with fewer than 20 individuals born there were excluded from this analysis.

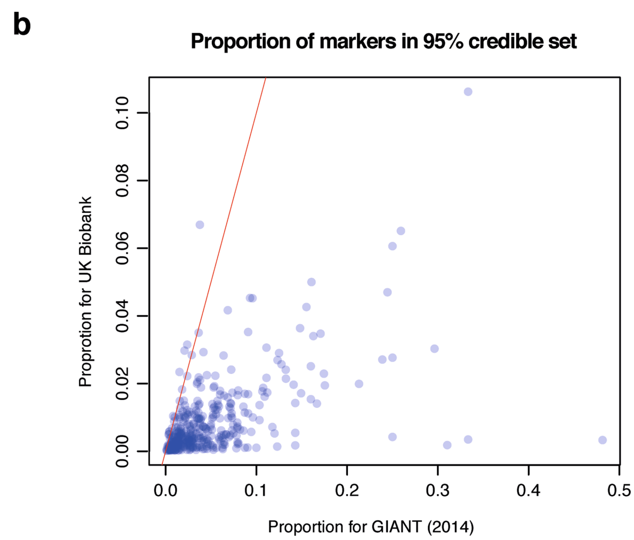


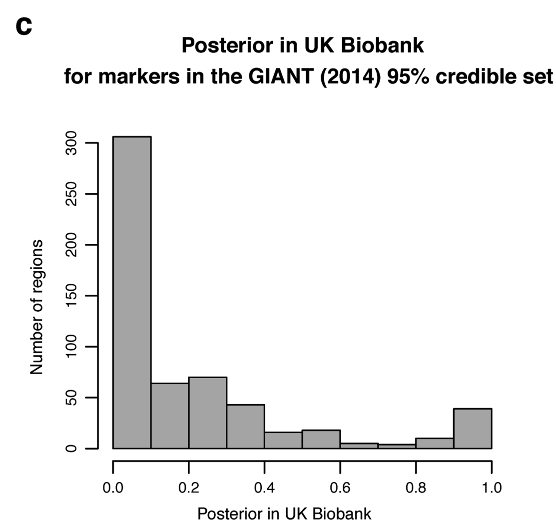
Extended Data Fig. 4 | Distribution of information scores at autosomal markers in the imputed dataset. The top left graph shows the full distribution of the information scores. The remaining panels show distributions in tranches of MAF; MAF > 5%, 1% ≤ MAF < 5%, 0.1% ≤ MAF < 1%, 0.01% ≤ MAF < 0.1% and 0.001% ≤ MAF < 0.01%.



Extended Data Fig. 5 | Example region of association in standing height GWAS. GWAS association statistics (*P* values) for standing height focusing on a ~3-Mb region of chromosome 2 that did not reach genome-wide significance in the GIANT (2014) meta-analysis, but did in UK Biobank (linear mixed model; see Methods). The *P* values shown are not adjusted for multiple testing. Markers genotyped in the UK Biobank are shown as diamonds, and imputed markers as circles. The two markers with the smallest *P* value for each of the genotyped data and imputed data are enlarged and highlighted with black outlines, and other UK Biobank markers are coloured according to their correlation (r^2) with one of these two. That is, genotyped markers with the leading genotyped marker (rs17713396), and imputed markers with the leading imputed marker (rs12714401). Markers with r^2 values of less than 0.1 are shown as black or green.







Extended Data Fig. 6 | Comparison of fine-mapping in GIANT (2014) and UK Biobank imputed data. Here we summarize results of our credible set analysis in GIANT (2014) and UK Biobank for 575 genomics regions associated with standing height in both studies (see Methods). A red solid line on a plot indicates where $x = y$. **a**, Both plots compare the number of markers in the 95% credible sets in which the size is less than 18 markers in both studies (363 regions in the left-hand plot; 445 in the right-hand plot). **b**, **c**, Both plots are from the analysis considering all markers in each study. In **b** we show, for each region, the proportion of markers used in the analysis for a given study that are in the 95% credible set for that study. The plot contains the same 363 regions as shown in the left-hand plot in **a**. In **c** we summarize, for all 575 regions, how much weight our UK Biobank analysis placed on markers that our analysis of GIANT (2014) indicated were important.

Extended Data Table 1 | Types and dates of data collection in UK Biobank

	Type of data	Date of data collection	Number of participants
			Anticipated
Questionnaire and interview	Sociodemographic data	Recruitment: 2006-2010[a]	500,000
	Family history and early life		500,000
	Psychosocial factors		500,000
	Lifestyle		500,000
	Medical history		500,000
	Cognitive function		500,000
Physical measures	Blood pressure	Recruitment: 2006-2010	500,000
	Hand grip strength		500,000
	Anthropometry		500,000
	Spirometry		500,000
	Heel bone density		500,000
	Arterial stiffness		200,000
	Hearing test		200,000
	Cardiorespiratory fitness plus ECG		100,000
	Eye measures		100,000
Web-based questionnaires	Diet	2011-2012	210,000[b]
	Cognitive function	2014	120,000
	Occupational history	2015	120,000
	Mental health	2016	150,000
	Irritable bowel syndrome	2017	150,000
Enhancements	Physical activity monitor	2013-2014[c]	100,000
	Biochemistry markers[d]	2006-2010	500,000
	Genotyping	2006-2010	500,000
	Multi-modal imaging[e]	2014-2022	100,000[f]
Electronic medical records	Death registry	2006-current	14,000
	Cancer registry	1971-current	79,000
	Hospital inpatient data	1996-current	400,000
	Primary care data	Birth-current	pending

[a]The baseline visit (including the touchscreen questionnaire, physical measures and biological sampling) was repeated approximately 5 years later (2012–2013) in a subset of 20,000 participants and in those who attended an imaging assessment centre (2014–2022).
[b]Includes 70,000 participants who completed the diet online questionnaire at the end of the recruitment visit[48].
[c]A repeat assessment of physical activity on four occasions over a 12-month period is being collected on 2,5000 of these participants (2018–2019).
[d]Biochemistry markers were measured in the baseline sample for 500,000 participants and in the repeat assessment sample for 20,000 participants. The urinary biomarkers were made available in 2016; the serum and red blood cell markers available are pending (at the time of press).
[e]The imaging study includes brain, heart and body MRI, carotid ultrasound and 12-lead ECG scan and a full-body dual energy X-ray absorptiometry scan, plus a repeat of the baseline assessment (including biological sampling). Repeat imaging in a subset of participants is expected to start in 2019.
[f]Data are currently available for 25,000 participants, with the remaining 75,000 participants to attend over the next few years. See Supplementary Table 1 for further information about these data types.

Extended Data Table 2 | The number of markers and samples by genotyping array at main stages of the UK Biobank genotyping experiment

		UK BiLEVE Axiom array only	UK Biobank Axiom array only	Both arrays	Total
Included in experiment	Number of samples sent to Affymetrix (including duplicates)	50561	443568	0	494078
Included in data delivery from Affymetrix	Number of markers	18019	34313	760096	812428
	Number of samples (including duplicates)	50520	438692	0	489212
Included in released data	Number of markers	17536	34197	753693	805426
	Number of unique samples	49950	438427	0	488377

'Data delivery from Affymetrix' refers to the data produced by Affymetrix after applying their filtering (Supplementary Information). 'Released data' refers to the released genotype data, after applying quality control measures, as detailed in sections S2 and S3 of the Supplementary Information.

Extended Data Table 3 | Counts and proportions of self-reported ethnic groups among 488,377 genotyped UK Biobank participants

Ethnic group	Self-reported ethnic background	Count of genotyped UK Biobank participants	
White			460,186 (94.23%)
	British	431,059 (88.26%)	
	Any other white background	15,821 (3.24%)	
	Irish	12,760 (2.61%)	
	White	546 (0.11%)	
Asian or Asian British			9,474 (1.94%)
	Indian	5,716 (1.17%)	
	Pakistani	1,748 (0.36%)	
	Any other Asian background	1,747 (0.36%)	
	Bangladeshi	221 (0.05%)	
	Asian or Asian British	42 (0.01%)	
Black or Black British			7,649 (1.57%)
	Caribbean	4,299 (0.88%)	
	African	3,206 (0.66%)	
	Any other Black background	118 (0.02%)	
	Black or Black British	26 (0.01%)	
Chinese			1,504 (0.31%)
	Chinese	1,504 (0.31%)	
Mixed			2,843 (0.58%)
	Any other mixed background	996 (0.2%)	
	White and Asian	802 (0.16%)	
	White and Black Caribbean	597 (0.12%)	
	White and Black African	402 (0.08%)	
	Mixed	46 (0.01%)	
Other/Unknown			6,721 (1.38%)
	Other ethnic group	4,357 (0.89%)	
	Not stated	2,364 (0.48%)	

Categories of self-reported ethnic background (UK Biobank data field 21000) and broader-level ethnic groups are shown here to reflect the two-layer branching structure of the ethnic background section in the UK Biobank touchscreen questionnaire[14]. Participants first picked one of the broader-level ethnic groups (for example, 'white'), and were then prompted to select one of the categories within that group (for example, 'Irish'). The broader-level groups are also shown here as an ethnic background category ('white' in column two) because a small proportion of participants only responded to the first question. In this table, we also combine the category 'other ethnic group' with an aggregated non-response category 'not stated', which includes all participants who did not know their ethnic group, or stated that they preferred not to answer, or did not answer the first question.

Extended Data Table 4 | Failure rates for six marker-based quality tests

Test	Average number of SNPs failed per batch (sd)	Fraction of all genotype calls affected
Affymetrix cluster QC	1109 (699)	0.00140
1. Batch effect	197 (86)	0.000249
2. Plate effect	284 (266)	0.000358
3. Departure from Hardy-Weinberg equilibrium	572 (77)	0.000723
4. Sex effect	45 (5)	0.0000569
5. Array effect[a]	5417	0.00683
6. Discordance across controls[b]	622 and 632	0.000796
Total	7704 (721)	0.00971

For all numbered tests, a marker (or marker within a batch) was set to missing if the test yielded $P < 10^{-12}$, except in the case of test 6, for which a marker was set to missing if the test yielded <95% concordance. See Supplementary Information for details of each test ($n = 463{,}844$ samples). The total is not equal to the sum of all tests because it is possible for a marker to fail more than one test. Because the two arrays contain slightly different sets of markers, the total number of genotype calls used to compute the fractions is:
$N_{\mathrm{ukbb}} L_{\mathrm{ukbb}} + N_{\mathrm{ukbl}} L_{\mathrm{ukbl}}$, in which N and L refer to the numbers of markers and samples typed on the UK Biobank Axiom array (ukbb) and samples typed on the UK BiLEVE Axiom array (ukbl) within the Affymetrix data delivery (see Supplementary Table 1).
[a]The array effect test was applied across all batches and only for markers present on both arrays, so we simply report the total number of markers that failed this test.
[b]The discordance test was applied across all batches, but not all markers are present on both arrays. The first value is the number of unique markers on the UK BiLEVE Axiom array that failed this test, and the second is for markers on the UK Biobank Axiom array.

Extended Data Table 5 | Summary of related pairs (third-degree relatives or closer) for the full UK Biobank cohort

	Monozygotic twins	Parent-offspring	Full siblings	2nd degree	3rd degree	Total
Number of pairs	179	6,276	22,666	11,113	66,928	107,162

Counts are derived from the kinship coefficients (see Methods). The count of monozygotic twins is after excluding samples identified as duplicates (Supplementary Information).

OPEN
https://doi.org/10.1038/s41586-018-0571-7

Genome-wide association studies of brain imaging phenotypes in UK Biobank

Lloyd T. Elliott[1], Kevin Sharp[1], Fidel Alfaro-Almagro[2], Sinan Shi[1], Karla L. Miller[2], Gwenaëlle Douaud[2], Jonathan Marchini[1,3,4]* & Stephen M. Smith[2,4]*

The genetic architecture of brain structure and function is largely unknown. To investigate this, we carried out genome-wide association studies of 3,144 functional and structural brain imaging phenotypes from UK Biobank (discovery dataset 8,428 subjects). Here we show that many of these phenotypes are heritable. We identify 148 clusters of associations between single nucleotide polymorphisms and imaging phenotypes that replicate at $P < 0.05$, when we would expect 21 to replicate by chance. Notable significant, interpretable associations include: iron transport and storage genes, related to magnetic susceptibility of subcortical brain tissue; extracellular matrix and epidermal growth factor genes, associated with white matter micro-structure and lesions; genes that regulate mid-line axon development, associated with organization of the pontine crossing tract; and overall 17 genes involved in development, pathway signalling and plasticity. Our results provide insights into the genetic architecture of the brain that are relevant to neurological and psychiatric disorders, brain development and ageing.

Brain structure and function vary between individuals and can be measured non-invasively using magnetic resonance imaging (MRI). The effects of neurological and psychiatric disorders such as Alzheimer's disease, Parkinson's disease, schizophrenia, bipolar disorder and autism can be seen in MRI data[1]. MRI can therefore provide intermediate endophenotypes that can be used to assess the genetic architecture of such disorders.

Structural MRI measures of brain anatomy include tissue and structure volumes, such as total grey matter volume and hippocampal volume, while other MRI modalities allow the mapping of different biological markers such as venous vasculature, microbleeds and aspects of white matter microstructure. Brain function is typically measured using task-based functional MRI (fMRI), in which subjects perform tasks or experience sensory stimuli; task-based fMRI uses imaging sensitive to local changes in blood oxygenation and flow caused by brain activity in grey matter. Brain connectivity can be divided into functional connectivity, where spontaneous temporal synchronizations between brain regions are measured using fMRI with subjects scanned at rest, and structural connectivity, measured using diffusion MRI (dMRI), which images the physical connections between brain regions based on how water molecules diffuse within white matter tracts. For those not familiar with the neuroimaging field, we have provided a glossary in Supplementary Note 1.

A new resource for relating neuroimaging to genetics is UK Biobank, a rich, long-term prospective epidemiological study of 500,000 volunteers[2]. Participants were 40–69 years old at recruitment, with one aim being to acquire as rich data as possible before disease onset. Identification of disease risk factors and early markers will increase over time with emerging clinical outcomes[3]. A brain and body imaging extension will scan 100,000 participants by 2020, with brain imaging including three structural modalities, resting and task-based fMRI, and diffusion MRI[4] (Supplementary Table 1). An automated image processing pipeline removes artefacts and renders images comparable across modalities and participants; it also generates thousands of image-derived phenotypes (IDPs), distinct measures of brain structure and function[5]. Example IDPs include the volume of grey matter in distinct brain regions, and measures of functional and structural connectivity between specific pairs of brain areas. The combination of large subject numbers with multimodal imaging data acquired using homogeneous hardware and software is a unique feature of UK Biobank.

Another key component of the UK Biobank resource has been the collection of genome-wide genetic data using a purpose-designed genotyping array. A custom quality control, phasing and imputation pipeline was developed to address the challenges specific to the experimental design, scale, and diversity of the UK Biobank dataset. The genetic data were publicly released in July 2017 and consist of about 96 million genetic variants in almost 500,000 participants[6].

Joint analysis of the genetic and brain imaging datasets produced by UK Biobank presents a unique opportunity for uncovering the genetic bases of brain structure and function, including genetic factors that are related to brain development, ageing and disease. In this study, we carried out genome-wide association studies (GWASs) for 3,144 IDPs, covering the entire brain and including 'multimodal' information on grey matter volume, area and thickness, white matter connections and functional connectivity, at 11,734,353 single-nucleotide polymorphisms (SNPs) in up to 8,428 individuals with both genetic and brain imaging data. We used two separate sets of data from UK Biobank to evaluate replication of significant genetic associations from the discovery phase. We also carried out multi-trait GWAS, SNP heritability analysis, genetic correlation analysis of IDPs with brain-related traits and an analysis of enrichment of genomic regions with different functions. Previous large-scale GWAS imaging studies have focused on narrower ranges of phenotypes including studies of: grey matter volume in seven subcortical regions by combining data across more than fifty studies[7,8]; whole-brain grey matter volumes and thicknesses by combining data from 59 acquisition sites[9]; and white matter connectivity in healthy young adult twins[10]. We expect that the homogeneous image acquisition and genetic data assay in UK Biobank will boost the power of our study.

[1]Department of Statistics, University of Oxford, Oxford, UK. [2]Centre for Functional MRI of the Brain (FMRIB), Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK. [3]The Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [4]These authors jointly supervised this work: Jonathan Marchini, Stephen Smith. *e-mail: marchini@stats.ox.ac.uk; steve@fmrib.ox.ac.uk

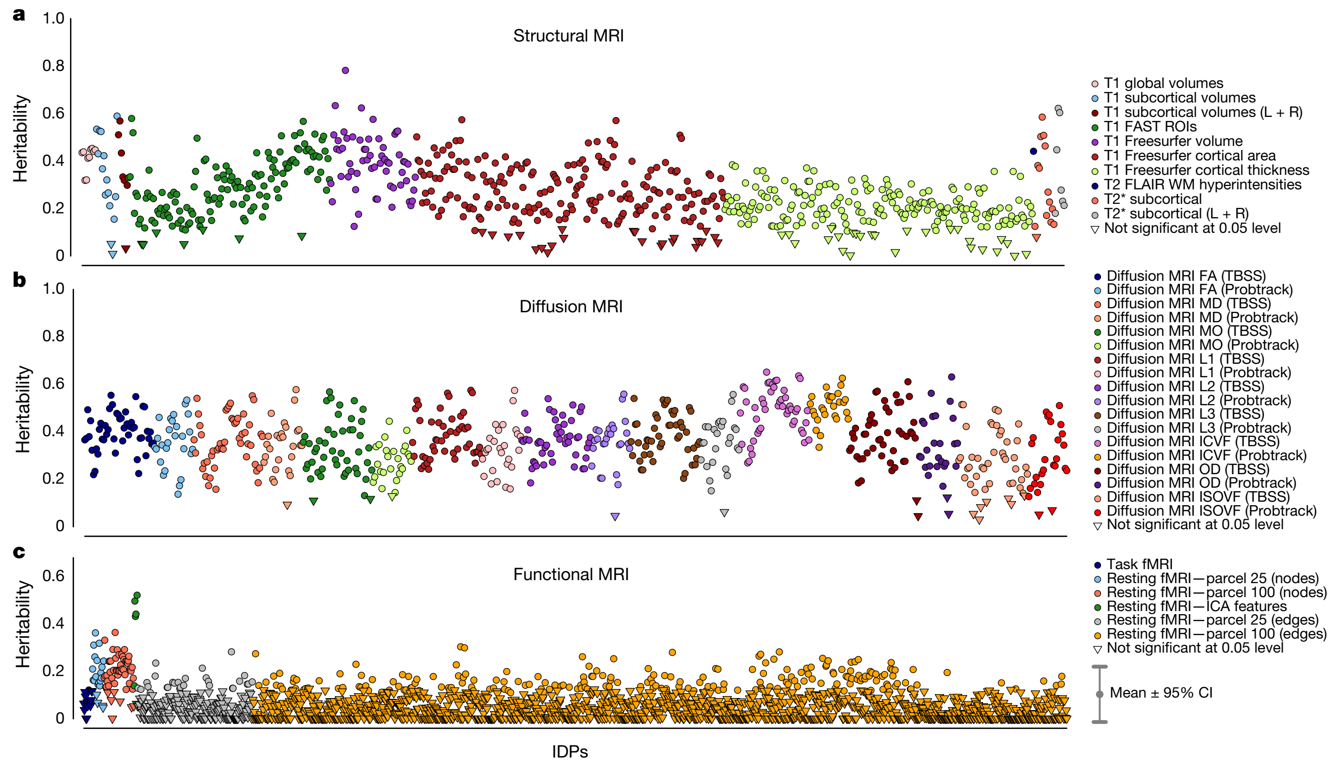


Fig. 1 | Estimated heritability of IDPs. Estimated heritability (*y*-axis) of all of the IDPs analysed ($n = 8{,}428$ subjects; see Methods for heritability calculation details). IDPs were split into three broad groups. **a**, Structural MRI. **b**, Diffusion MRI. **c**, Functional MRI. Points are coloured according to IDP groups. Circles and inverted triangles, respectively, are used to identify IDPs that do and do not have heritability significantly different from 0 at the 5% significance level. The mean 95% confidence interval (CI) error bar size is indicated at the bottom right.

The UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) to obtain and disseminate data and samples from the participants (http://www.ukbiobank.ac.uk/ethics/), and these ethical regulations cover the work in this study. Written informed consent was obtained from all participants.

All results are available on the Oxford Brain Imaging Genetics (BIG) web browser (http://big.stats.ox.ac.uk/), which allows users to browse associations by SNP, gene or phenotype. This was built from the PheWeb code base (https://github.com/statgen/pheweb/) and extended to allow easier searching of phenotypes. In addition to the brain IDP GWAS results, the browser also includes GWAS results from more than 2,500 other traits and diseases.

Heritability and genetic correlations of IDPs

Figure 1 shows the estimated SNP heritability (h^2) of all IDPs and whether h^2 is significantly different from 0 at the nominal 5% significance level (Supplementary Table 2, Supplementary Fig. 1). Out of 3,144 IDPs, 1,578 show significant SNP heritability. Of the structural MRI IDPs, volumetric measures are the most heritable and cortical thicknesses the least. Of the diffusion MRI measures, the tractography-based IDPs show lower heritability than the tract-skeleton-based IDPs. The resting-state fMRI functional connectivity edges show the lowest levels of SNP heritability, with just 235 of 1,771 IDPs being significant, which is consistent with additive heritability estimates from twin studies of network edges from fMRI and magnetoencephalography in the Human Connectome Project[11]. However, four of the six resting fMRI features identified by independent component analysis (ICA; estimated as data-driven reductions of this full set of fMRI edges) are much more highly heritable. By contrast, most of the resting-state node amplitude IDPs show significant evidence of SNP heritability; the task-related fMRI IDPs do not.

We found lower levels of SNP heritability for subcortical volumes than previously estimated in twin studies[12–14] (Supplementary Fig. 2). This is typical of many traits in the literature[15] and may result from upward bias in twin study estimates due to gene–gene and gene–environment interactions[16,17], or downward bias of SNP heritability due to uncaptured rare genetic variation. We also compared the GWAS results for seven subcortical volumes with those obtained by the ENIGMA consortium (http://enigma.ini.usc.edu/research/download-enigma-gwas-results/), via a genetic correlation analysis (Supplementary Table 3). There was a strong correlation between the studies, suggesting that there were no major differences in how these phenotypes were measured. In all cases a perfect genetic correlation of 1 lies within the 95% confidence intervals.

Supplementary Fig. 3 shows the genetic correlations, together with the raw phenotype correlations, for several groups of analysed IDPs. There is a range of both strong and weak, positive and negative genetic correlations between the IDPs.

Significant associations between IDPs and SNPs

In all analyses we estimated genetic effects with respect to the number of copies of the non-reference allele. Using a minor allele frequency filter of 1% and a $-\log_{10}(P\text{ value})$ threshold of 7.5, we found 1,262 significant associations between SNPs and the 3,144 IDPs. These associations spanned all classes of IDPs, except task-related fMRI (Supplementary Table 4), with the swMRI T2* group showing a relatively large number of associations. The $-\log_{10}(P\text{ value})$ threshold of 7.5 controls for the number of tests carried out across SNPs and accounts for the correlation structure between genetic variants. Of these 1,262 associations, 844 and 455 replicated at the 5% significance level using our two smaller replication datasets (see Methods and Supplementary Table 5). Some associated genetic loci overlapped across IDPs; we estimate that there are approximately 427 distinct associated genetic regions (clusters). One hundred and forty-eight of these clusters have a lead SNP that replicates at the 5% level in our replication set of 3,456 participants, and 91 below a 5% false discovery rate (FDR) threshold. We would expect about 21 of the lead SNPs in the 148 clusters to replicate under a null hypothesis of no association.

At a threshold of $-\log_{10}(P) > 11$, which additionally corrects for all 3,144 GWAS carried out (see Methods), we found 368 significant associations between genetic regions and distinct IDPs (Supplementary Table 6, Supplementary Fig. 4). These associations with 78 unique SNPs can be grouped together into 38 distinct clusters by grouping across IDPs (Extended Data Table 1). Taking our lead SNP in each of the 38 regions, we found that all 38 had $P < 0.05$ in our replication set of 3,456 participants, and all 38 were significant at 5% FDR. We found no appreciable change in these GWAS results when we included a set of potential body confound measures in addition to the main set of imaging confound measures (see Methods and Supplementary Fig. 5). We also carried out a winner's curse corrected post-hoc power analysis that agreed well with the results of our replication studies (Supplementary Note 2).

Supplementary Figs. 6 and 7 provide genome-wide association plots (also known as Manhattan plots) and QQ-plots for all 3,144 IDPs and the subset of IDPs listed in Extended Data Table 1, respectively. Having identified a SNP as being associated with a given IDP, it can be useful then to explore the association with all other IDPs via a PheWAS (phenome-wide association study) plot. Supplementary Fig. 8 shows the PheWAS plots for all 78 SNPs listed in Supplementary Table 6 with $-\log_{10}(P) > 11$. The Oxford Brain Imaging Genetics (BIG) web browser (http://big.stats.ox.ac.uk/) allows researchers to view the PheWAS for any SNP of interest. We found that 4 of the 78 SNPs were associated ($P < 0.05/3{,}144$; that is, $-\log_{10}(P) > 4.79$) with all 3 classes of structural, dMRI and functional measures, and these were all SNPs in cluster 31 of Extended Data Table 1 (Supplementary Fig. 8, pages 62–65). This genetic locus is associated with the volume of the precuneus and cuneus, dMRI measures for the forceps major (a fibre bundle that connects the left and right cuneus), and two functional connections (parcellation 100 edges 614 and 619, which connect the precuneus to other cognitive networks). Supplementary Fig. 9 illustrates the sharing of association signal across IDPs at the 615 unique SNPs listed in Supplementary Table 5. Supplementary Fig. 10 shows the relationship between the number of associations found and the estimated SNP heritability for each IDP.

Overall, our results clearly replicate the majority of the loci identified by the ENIGMA consortium in two separate GWASs of seven brain subcortical volume IDPs in up to 13,171 subjects[7], and of hippocampal volume in 33,536 subjects (although not all reached genome-wide significance, probably owing to the smaller sample size in our study; Supplementary Fig. 11). We also replicate an association between volume of white matter hyperintensities ('lesions') and SNPs in *TRIM47* (for example, rs3744017, $P = 1.4 \times 10^{-12}$, cluster 37)[18].

It can be challenging to interpret precisely the function of SNPs identified in a GWAS. Most of the SNPs in the 38 loci in Extended Data Table 1 are either in genes, including 7 missense SNPs and 2 SNPs in untranslated regions (UTRs), or in high linkage disequilibrium with SNPs that are themselves in the genes of interest, and many are significant expression quantitative trait loci (eQTLs) in the GTEx database[19]. In total, we found 17 genetic loci that can be linked to genes that broadly contribute to brain development, patterning and plasticity (out of the 38 clusters reported in Extended Data Table 1; for more details, see Supplementary Note 3). Below we focus on some of the most compelling examples.

A major source of cross-subject differences seen in T2* data are microscopic variations in magnetic field, often associated with iron deposition in ageing and pathology[20]. We identified many associations between T2* in the caudate nucleus, putamen and pallidum and SNPs in genes (*TF*, rs4428180, $P = 2.23 \times 10^{-22}$; *HFE*, rs1800562, $P = 6.6 \times 10^{-20}$; *SLC25A37*, rs35469695, $P = 2.22 \times 10^{-12}$) or near genes (*FTH1*, rs11230859, $P = 2.31 \times 10^{-17}$) that are known to affect iron transport and storage, or neurodegeneration with brain iron accumulation (NBIA)[21] (*COASY*, rs668799, $P = 1.43 \times 10^{-17}$). In addition, we identified four SNPs that either encode or are eQTLs of genes involved in transport of nutrients and minerals: *SLC44A5* (rs76934732, $P = 8.51 \times 10^{-13}$), *SLC39A8* (also known as *ZIP8*; rs13107325, $P = 1.04 \times 10^{-42}$), *SLC20A2* (rs2923405, $P = 3.31 \times 10^{-17}$) and *SLC39A12* (also known as *ZIP12*; rs10764176, $P = 3.3 \times 10^{-21}$). For more details, see Supplementary Note 3.

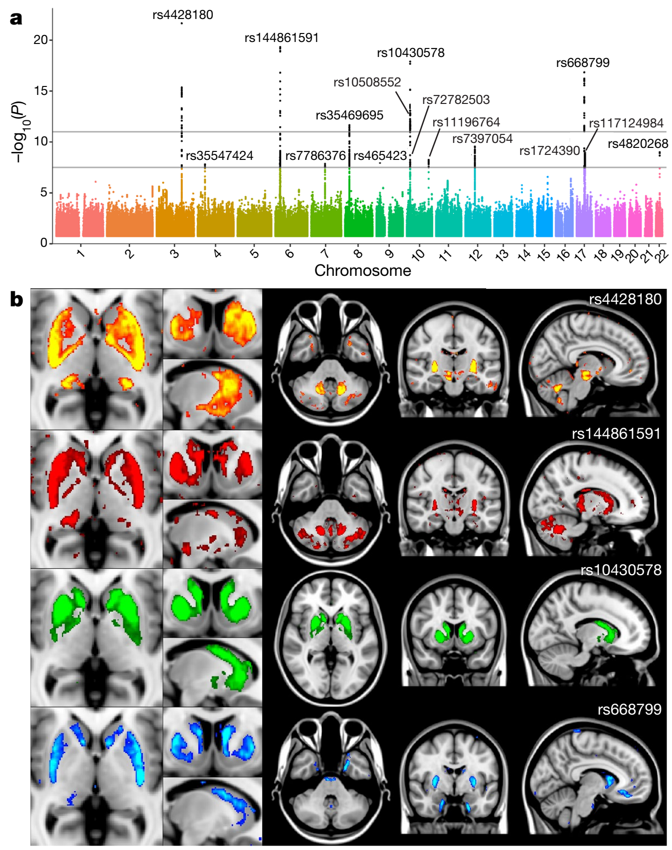


Fig. 2 | Manhattan plot and spatial mapping of the associations between T2* in the putamen and four SNPs. a, The Manhattan plot relates to the original GWAS for the IDP T2* in the bilateral putamen. The lower grey line indicates the $-\log_{10}(P$ value) threshold of 7.5 and the upper line the threshold of 11 (see main text). **b**, The spatial maps show that the four SNPs (one per row) most strongly associated with T2* in the putamen have distinct voxelwise patterns of effect across the whole brain: the effect of rs4428180 (*TF*) is found in the dorsal putamen and body of the caudate nucleus, but also in the right subthalamic nucleus and substantia nigra, red nucleus, lateral geniculate nucleus of the thalamus and dentate nucleus; rs144861591 (*HFE*) in the dorsal striatum, subthalamic nucleus, dentate nucleus and Crus I/II of the cerebellum; rs10430578 (*SLC39A12*) in the whole dorsal striatum and pallidum; and rs668799 (*COASY*) in the whole dorsal striatum, subgenual cingulate cortex and entorhinal cortex. The standard MNI152 T1 image is used as background for the spatial maps (left is right). All group difference images (colour overlays) are thresholded at a T2* difference of 0.6 ms. These voxelwise SNP association maps were calculated from the discovery sample of 8,428 subjects (see main text).

Interrogating images at a voxel-wise level can provide further insight about the detailed spatial localization of SNP associations and can possibly identify additional associated areas not already well captured by IDPs (while keeping in mind the statistical dangers of potential circularity[22]). For instance, by looking at the difference between the average T2* image from subjects with no copies versus one copy of the rs4428180 (*TF*) non-reference allele, we found effects of this SNP not just in the putamen and pallidum, but also in additional, smaller regions of subcortical structures not included as IDPs (Fig. 2). We similarly created in Fig. 2 the voxelwise differences associated with three additional SNPs, from the most significant GWAS associations with T2* in the putamen as seen in the Manhattan plot. This approach also allowed us to observe grey matter volume effects across the entire brain associated with rs13107325 (*SLC39A8*; Extended Data Fig. 1), which has been linked in previous (non-imaging) GWASs to intelligence[23], schizophrenia[24], blood pressure[25] and higher risk of cardiovascular

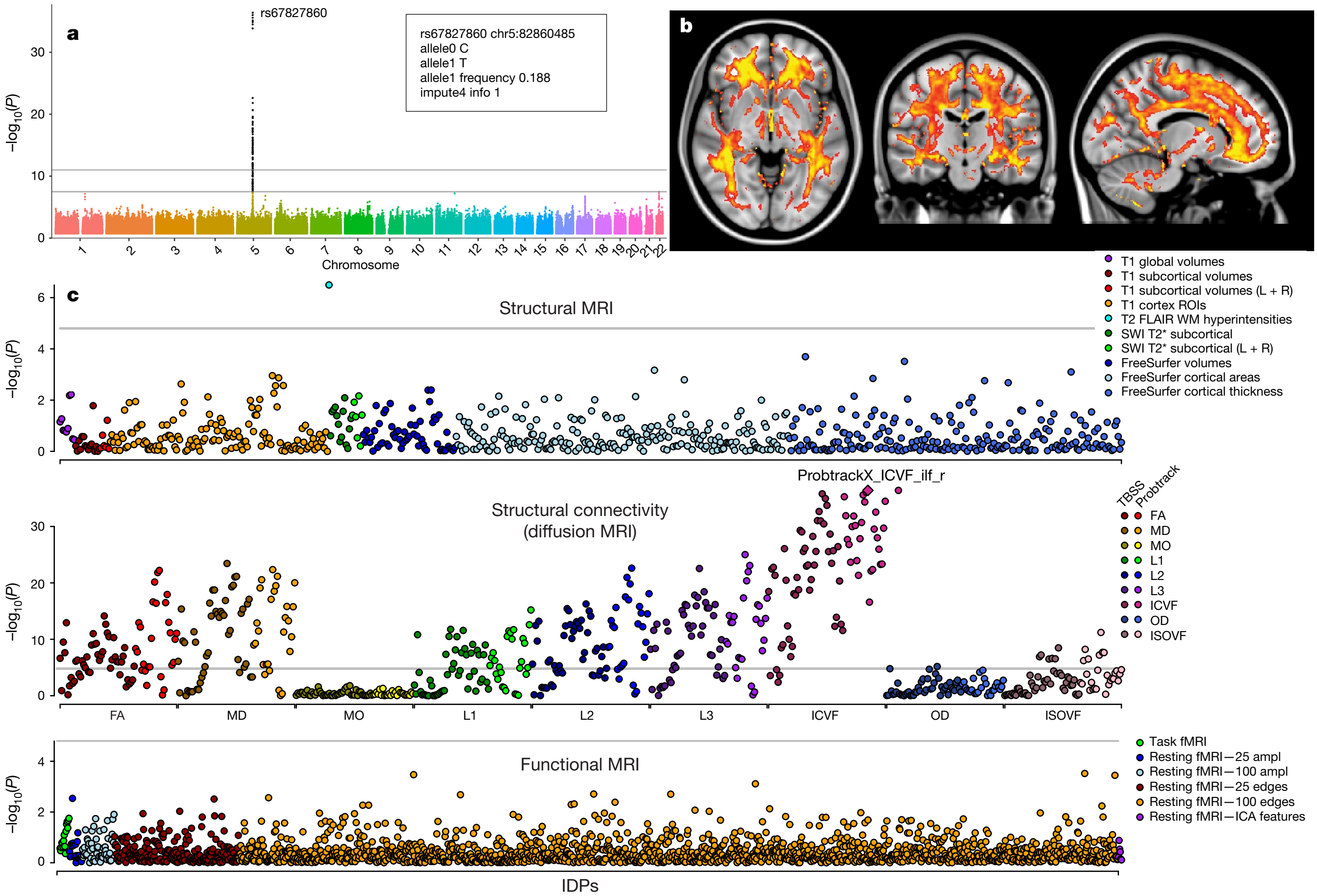


Fig. 3 | Manhattan plot, spatial mapping and PheWAS plot relating to the association between the dMRI ICVF measure and rs67827860 (*VCAN*). a, The Manhattan plot relates to the original IDP GWAS with the strongest association (ICVF in the right inferior longitudinal fasciculus using tractography, associated with rs67827860). The ICVF parameter, estimated from the NODDI modelling[36], aims to quantify predominantly intra-axonal water in white matter, by estimating where water diffusion is restricted. Summary details of SNP rs67827860 are given in the top right box. The lower grey line indicates the $-\log_{10}(P \text{ value})$ threshold of 7.5 and the upper line the threshold of 11. **b**, Spatial mapping of rs67827860 against voxelwise ICVF in white matter (ICVF was averaged across all 4,957 subjects with zero copies of the non-reference allele, and the average from all 2,304 subjects that had one copy was subtracted from that, for display in colour here; the difference was thresholded at 0.005 (unitless fractional measure)). Unlike the examples of (spatially) very focal effects in T2* and grey matter volume in Fig. 2 and Extended Data Fig. 1, the effects of this SNP are extremely widespread across most of the white matter tracts (associated with 45 out of the 199 IDPs in cluster 11, Supplementary Table 5). **c**, The PheWAS plot for SNP rs67827860 shows the association ($-\log_{10}(P)$) on the y-axis for the SNP with each of the 3,144 IDPs. The IDPs are arranged on the x-axis in the three panels: structural MRI IDPs (top), dMRI IDPs (middle) and fMRI IDPs (bottom). Points are coloured to delineate subgroups of IDPs. Grey lines show the Bonferroni multiple testing threshold of 4.79. In addition to the IDP of white matter hyperintensities volume, there is a notable association with numerous dMRI IDPs (especially diffusion tensor-derived measures of fractional anisotropy, mean diffusivity and L1, L2 and L3 eigenvalues of the diffusion tensor, as well as additional ICVF measures).

death[26]. These effects could now be observed in a relevant brain region, the anterior cingulate cortex, which has multifaceted roles including in fluid intelligence[27], schizophrenia[28] and modulating autonomic states of cardiovascular arousal[29].

Notably, three SNPs related to our white matter IDPs were in genes or eQTLs of genes encoding three proteins of the extracellular matrix (ECM): rs2365715 ($P = 5.38 \times 10^{-12}$), an eQTL of *BCAN*, is associated with one dMRI microstructural measure in the genu of the corpus callosum; rs3762515 ($P = 4.27 \times 10^{-13}$), in the 5′ UTR of *EFEMP1*, with the volume of white matter lesions; and rs67827860 ($P = 4.06 \times 10^{-37}$, Fig. 3), located in an intron of *VCAN*, with multiple dMRI measures of most white matter tracts (199 IDPs in total). Overall, the vast majority of forebrain white matter-related dMRI IDPs were associated with SNPs related to genes that encode proteins involved in the extracellular matrix and epidermal growth factor signalling. These proteins have key roles in synaptic plasticity and myelin repair, and are associated with multiple sclerosis, stroke, amyotrophic lateral sclerosis and major depressive disorder (Supplementary Note 3).

Two additional examples further illustrate meaningful correspondences between the locations of our brain IDPs and significantly associated genes. First, the volume of the fourth ventricle, which develops from the central cavity of the neural tube, was found to be significantly associated with a SNP in, and eQTL of, *ALDH1A2* (rs2642636, $P = 5.2 \times 10^{-16}$). This gene encodes an enzyme that facilitates posterior organ development and prevents human neural tube defects, including spina bifida[30]. Second, we found two SNPs associated with dMRI IDPs of the crossing pontine tract (the part of the pontocerebellar fibre bundle that arises from the pontine nuclei and decussates across the brain midline to project to the contralateral cerebellar cortex) in genes that regulate axon guidance and fasciculation during development (*SEMA3D*, rs2286184, $P = 5.31 \times 10^{-17}$ and *ROBO3*, rs4935898 (missense), $P = 1.76 \times 10^{-19}$; Fig. 4). The exact location of our IDP in the crossing fibres of the pons coincides with the function of *ROBO3*, which is specifically required for axons to cross the midline in the hindbrain (pons, medulla oblongata and cerebellum); mutations in *ROBO3* result in horizontal gaze palsy, a disorder in which the corticospinal and

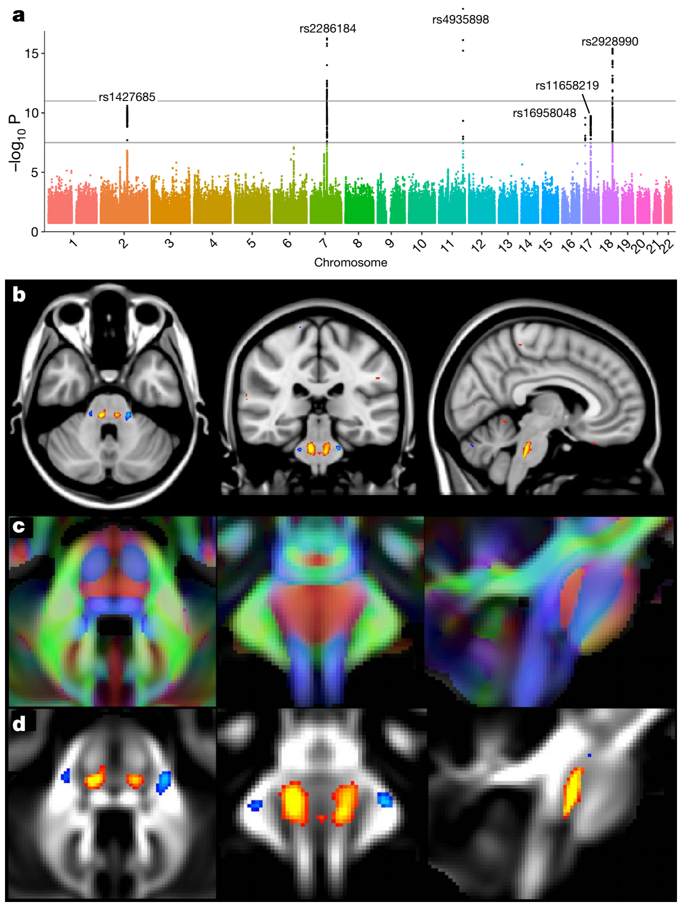


Fig. 4 | Manhattan plot and spatial mapping of the association between the dMRI tensor mode measure and SNP rs4935898 (*ROBO3*). a, The Manhattan plot relates to the original GWAS for the IDP of tensor mode in the crossing pontine tract associated with rs4935898. **b–d**, Tensor mode was averaged across all 6,807 subjects with approximately zero copies of the non-reference allele, and the average from all 703 subjects that had approximately one copy was subtracted from that, for display in red/yellow–blue/light blue here, thresholded at 0.05 (**b**, **d**). **b**, Results are shown overlaid on the MNI152 T1 structural image; by contrast, background in **c** and **d** is the UK Biobank average fractional anisotropy image, which shows clear tract structure within the brainstem. **c**, Orientation of the fibre tracts (in red, running left to right). The spatial distribution (not shown) for the effects of rs2286184 (*SEMA3D*) on tensor mode is almost identical to that of rs4935898, being again extremely spatially specific, with no extended effect elsewhere in the brain. These voxelwise SNP association maps were calculated from the discovery sample of 8,428 subjects (see main text).

somatosensory axons fail to cross the midline in the medulla[31]. Notably, all three significant associations with the IDP of the crossing pontine tract were found using the tensor mode of anisotropy (MO), a measure that is particularly useful in regions of crossing fibres[32].

Multi-phenotype association tests

One alternative strategy for analysing large numbers of IDPs is to use multi-trait tests that fit joint models of associations to groups of IDPs. Such approaches can use estimates of genetic correlation to boost power. In addition, by analysing P traits in one GWAS, these tests can avoid the need to correct for multiple genome-wide scans. We used a multi-trait test (see Methods) to analyse 23 groups of IDPs with up to 243 IDPs per group. These IDP groups were chosen to cover the majority of the IDP classes with significant IDP correlations in each grouping (Supplementary Table 7). Supplementary Fig. 12 shows the Manhattan plots for these genome-wide scans. Overall, across these 23 groups, we found 278 SNPs at about 160 loci associated with $-\log_{10}(P) > 7.5$ (Supplementary Table 8). Of these 278 SNPs, 170 survived a correction for 23 scans with $-\log_{10}(P) > 8.86$ and 138 of these 170 SNPs had a P value < 0.05 in the larger replication set of 3,456 samples. There can be large differences in P values between the multi-trait tests and the individual IDP tests (Supplementary Fig. 13), especially when taking account of the smaller number of tests carried out by the multi-trait approach (Supplementary Fig. 14). We found 25 loci that showed both a significant and replicated multi-trait association for an IDP group, while showing no genome-wide significance in the flanking region for any individual IDP in the corresponding group (Supplementary Table 9, Supplementary Note 3).

Three of these loci showed associations with the dMRI tensor mode of anisotropy measures (rs62073157, $P = 4.07 \times 10^{-11}$; rs35884657, $P = 1.04 \times 10^{-9}$; rs9939914, $P = 1.15 \times 10^{-11}$) and all were eQTLs of microtubule-related genes (*MAPT*, *TUBA1B* and *TUBB3*, respectively). The extended *MAPT* region has been repeatedly associated with Alzheimer's and Parkinson's diseases, frontotemporal dementia, and progressive supranuclear palsy (Supplementary Note 3).

Another example of the value of multi-trait testing can be seen in the association between IDPs of global brain volume measurements and an SNP located between *BANK1* and *SLC39A8*, which was previously identified in a GWAS of schizophrenia[33] (rs35518360, $P = 4.07 \times 10^{-12}$). This locus is also part of a multimodal cluster from our single-trait GWAS that includes subcortical and cerebellar grey matter volumes, pallidum T2* and dMRI in midbrain white matter tracts (cluster 10 in Supplementary Table 6). The multi-trait test thus made it possible to uncover this additional association between global brain volume measurement and this locus, which might prove relevant for better understanding observations of smaller brain volume in (particularly first episode or drug-naive) patients with schizophrenia[34].

Genetic correlation with clinically relevant traits

We measured the genetic correlation between a subset of heritable IDPs and ten neurodegenerative, psychiatric and personality traits (see Methods). We found suggestive evidence of genetic correlation for amyotrophic lateral sclerosis (ALS), schizophrenia and stroke, mainly with dMRI measures in white matter tracts (Supplementary Fig. 15). Supplementary Table 10 contains genetic correlation estimates for all IDP–trait combinations; see Supplementary Note 4 for further details.

Partitioning heritability by functional annotation

We applied a statistical approach that partitions the additive genetic heritability of a set of common variants for each of the 3,144 IDPs according to 24 functional annotations of the genome[35]. Extended Data Fig. 2 summarizes which functional annotations show enrichment stratified by 23 groups of IDPs (Supplementary Table 11). We find that regions of the genome annotated as super enhancers and several histone modifications show enrichment across many of the structural and diffusion IDP groups. Regions of the genome enriched for trimethylation of lysine 27 on histone H3 (H3K27me3) (and indicating strong evidence for silenced genes) show depletion of heritability across many of the IDP classes (Supplementary Fig. 16). IDP groups such as T1 subcortical volumes, dMRI fractional anisotropy (FA) and intracellular volume fraction (ICVF) show the strongest evidence of enrichment across multiple categories. The resting fMRI connectivity edge IDPs show no elevated enrichment, consistent with these traits showing low heritability (Fig. 1). Supplementary Fig. 17 shows this partitioning analysis for each IDP.

Conclusions

Bringing together researchers with backgrounds in brain imaging and genetic association was key to this work. We have uncovered a large number of associations at the nominal level of GWAS significance ($-\log_{10}(P) > 7.5$) and at a more stringent threshold ($-\log_{10}(P) > 11$) designed to (probably over-conservatively) control for the number of IDPs tested. Our use of multi-trait tests uncovered further novel loci. We find associations with all the main IDP groups except the task fMRI measures (despite these measures containing usable signal, for example having unique cognitive associations[4]).

We mainly found associations between MRI measures and genes involved in brain development and plasticity, as well as genes contributing to the transport of iron, nutrients and minerals (Supplementary Note 3). The genes linked to brain development and plasticity tended to be related to mental health disorders, including major depression disorder and schizophrenia, whereas those that encoded iron-related proteins tended to be related to neurodegenerative disorders, such as amyotrophic lateral sclerosis, Parkinson's disease and Alzheimer's disease. We also uncovered enrichments of functional annotations for many of the structural and diffusion IDPs.

A valuable aspect of this work has been to link the associated SNPs back to spatial properties of the voxel-level brain imaging data. For example, we have linked SNPs associated with IDPs to both highly spatially localized and widely spatially distributed effects, restricting these voxelwise analyses to the same imaging modality from which the original phenotypic association was found (though of course other modalities could also be tested in the same way). In addition, looking at PheWAS plots has been useful when working with so many phenotypes. It has allowed us to investigate the overall patterns of association and has led to the identification of SNP associations that span multiple modalities.

We used two additional sets of 930 and 3,456 samples to replicate a large number of the associations uncovered at the discovery phase. Over coming years, the number of UK Biobank participants for whom imaging data are available will increase to 100,000, allowing more complete discovery of the genetic basis of human brain structure, function and connectivity. Combining the discovery and replication samples is also likely to lead to novel associations, as will the use of methods that can analyse the huge IDP × SNP matrix of summary statistics of association. A potential avenue of research will involve attempts to uncover causal pathways that link genetic variants to IDPs and then to a range of neurological, psychiatric and developmental disorders.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0571-7.

Received: 3 November 2017; Accepted: 4 September 2018;
Published online 10 October 2018.

1. Toga, A. W. *Brain Mapping: An Encyclopedic Reference* (Academic, Amsterdam, 2015).
2. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. **12**, e1001779 (2015).
3. Allen, N. et al. UK Biobank: Current status and what it means for epidemiology. *Health Policy Technol*. **1**, 123–126 (2012).
4. Miller, K. L. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci*. **19**, 1523–1536 (2016).
5. Alfaro-Almagro, F. et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400–424 (2018).
6. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* https://doi.org/10.1038/s41586-018-0579-z (2018).
7. Hibar, D. P. et al. Common genetic variants influence human subcortical brain structures. *Nature* **520**, 224–229 (2015).
8. Hibar, D. P. et al. Novel genetic loci associated with hippocampal volume. *Nat. Commun*. **8**, 13624 (2017).
9. Shen, L. et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *Neuroimage* **53**, 1051–1063 (2010).
10. Koran, M. E. et al. Impact of family structure and common environment on heritability estimation for neuroimaging genetics studies using sequential oligogenic linkage analysis routines. *J. Med. Imaging (Bellingham)* **1**, 014005 (2014).
11. Colclough, G. L. et al. The heritability of multi-modal connectivity in human brain activity. *eLife* **6**, e20178 (2017).
12. Roalf, D. R. et al. Heritability of subcortical and limbic brain volume and shape in multiplex-multigenerational families with schizophrenia. *Biol. Psychiatry* **77**, 137–146 (2015).
13. den Braber, A. et al. Heritability of subcortical brain measures: a perspective for future genome-wide association studies. *Neuroimage* **83**, 98–102 (2013).
14. Kremen, W. S. et al. Genetic and environmental influences on the size of specific brain regions in midlife: the VETSA MRI study. *Neuroimage* **49**, 1213–1223 (2010).
15. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet*. **47**, 1114–1120 (2015).
16. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
17. Purcell, S. Variance components models for gene-environment interaction in twin analysis. *Twin Res*. **5**, 554–571 (2002).
18. Fornage, M. et al. Genome-wide association studies of cerebral white matter lesion burden: the CHARGE consortium. *Ann. Neurol*. **69**, 928–939 (2011).
19. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
20. Duyn, J. MR susceptibility imaging. *J. Magn. Reson*. **229**, 198–207 (2013).
21. Dusi, S. et al. Exome sequence reveals mutations in CoA synthase as a cause of neurodegeneration with brain iron accumulation. *Am. J. Hum. Genet*. **94**, 11–22 (2014).
22. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci*. **4**, 274–290 (2009).
23. Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet*. **50**, 912–919 (2018).
24. Goes, F. S. et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. B. Neuropsychiatr. Genet*. **168**, 649–659 (2015).
25. International Consortium for Blood Pressure Genome-Wide Association Studies Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).
26. Johansson, Å. et al. Genome-wide association and Mendelian randomization study of NT-proBNP in patients with acute coronary syndrome. *Hum. Mol. Genet*. **25**, 1447–1456 (2016).
27. Duncan, J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci*. **14**, 172–179 (2010).
28. Dolan, R. J. et al. Dopaminergic modulation of impaired cognitive activation in the anterior cingulate cortex in schizophrenia. *Nature* **378**, 180–182 (1995).
29. Critchley, H. D. et al. Human cingulate cortex and autonomic control: converging neuroimaging and clinical evidence. *Brain* **126**, 2139–2152 (2003).
30. Deak, K. L. et al. Analysis of ALDH1A2, CYP26A1, CYP26B1, CRABP1, and CRABP2 in human neural tube defects suggests a possible association with alleles in ALDH1A2. *Birth Defects Res. A Clin. Mol. Teratol*. **73**, 868–875 (2005).
31. Jen, J. C. et al. Mutations in a human *ROBO* gene disrupt hindbrain axon pathway crossing and morphogenesis. *Science* **304**, 1509–1513 (2004).
32. Douaud, G. et al. DTI measures in crossing-fibre areas: increased diffusion anisotropy reveals early white matter alteration in MCI and mild Alzheimer's disease. *Neuroimage* **55**, 880–890 (2011).
33. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
34. Haijma, S. V. et al. Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophr. Bull*. **39**, 1129–1138 (2013).
35. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet*. **47**, 1228–1235 (2015).
36. Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. NODDI: practical *in vivo* neurite orientation dispersion and density imaging of the human brain. *Neuroimage* **61**, 1000–1016 (2012).

Acknowledgements The data used in this work were obtained from UK Biobank (Data Application 8107). We are grateful to UK Biobank for making the data available, and to all UK Biobank study participants, who generously donated their time to make this resource possible. F.A.-A. acknowledges funding from the UK Medical Research Council and the Wellcome Trust via UK Biobank. K.L.M. and S.M.S. receive further support from the Wellcome Trust. J.M. acknowledges funding for this work from the European Research Council (ERC; grant 617306) and the Leverhulme Trust. G.D. acknowledges funding from the Medical Research Council UK (MR/K006673/1). We thank B. Fischl, D. Greve and M. Glasser for advice on FreeSurfer processing; J. Diprose and R. Esnouf for support with high-performance computing; S. McRobert for help with the Oxford BIG browser; and T. Nichols and A. Winkler for discussions about imaging confounds. For the genetic correlation analysis we used summary statistical data from several GWASs of brain-related conditions as follows: the ISGC Cerebrovascular Disease Knowledge Portal, International Genomics of Alzheimer's Project (IGAP), the Project MinE GWAS Consortium, the Social Science Genetic Association Consortium (SSGAC), the University of Exeter research group on Type 2 Diabetes, Obesity, Growth & Reproductive Ageing Genetics, the Psychiatric Genomics Consortium (PGC) and the ENIGMA consortium. We thank these groups for making these data publicly available and all the participants in these studies.

Author contributions J.M. and S.M.S. conceived and supervised the work. F.A.-A., K.L.M., G.D. and S.M.S. created the IDPs and imaging confounds. L.T.E., K.S., S.S. and J.M. carried out the genetic association, heritability, genetic correlation and functional enrichment analysis and created the Oxford BIG browser. J.M., S.M.S., G.D., F.A.-A., K.M., K.S. and L.T.E. interpreted the results and wrote the paper.

Competing interests J.M. is a co-founder and director of GENSCI Ltd. S.M.S. is a co-founder of SBGneuro. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0571-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0571-7.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.M. or S.M.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Imaging data and derived phenotypes. The UK Biobank brain imaging protocol consists of six distinct modalities covering structural, diffusion and functional imaging, summarized in Supplementary Table 1. For this study, we primarily used data from the February 2017 release of ~10,000 participants' imaging data (and an additional ~5,000 subjects' data released in January 2018 provided the larger replication sample).

The raw data from these six modalities have been processed for UK Biobank to create a set of IDPs[4,5]. These are available from UK Biobank, and it is these IDPs from the 2017–2018 data releases that we used in this study.

In addition to the IDPs directly available from UK Biobank, we created two extra sets of IDPs. First, we used FreeSurfer v6.0.0[37,38] (https://surfer.nmr.mgh.harvard.edu) to model the cortical surface (inner and outer 2D surfaces of cortical grey matter), as well as modelling several subcortical structures. We used both the T1 and T2 FLAIR images as inputs to the FreeSurfer modelling (or just the T1 when the T2 was not available). FreeSurfer estimates a large number of structural phenotypes, including volumes of subcortical structures, surface area of parcels identified on the cortical surface, and grey matter cortical thickness within these areas. The areas are defined by mapping an atlas containing a canonical cortical parcellation onto an individual subject's cortical surface model, thus achieving a parcellation of that surface. Here we used two atlases in common use with FreeSurfer: the Desikan-Killiany–Tourville atlas (denoted DKT[39]) and the Destrieux atlas (denoted a2009s[40]). The DKT parcellation is gyrus-based, whereas Destrieux aims to model both gyri and sulci based on the curvature of the surface. Cortical thickness is averaged across each parcel from each atlas, and the cortical area of each parcel is estimated, to create two IDPs for each parcel. Finally, subcortical volumes are estimated, to create a set of volumetric IDPs.

Second, we applied a dimension reduction approach to the large number of functional connectivity IDPs. Functional connectivity IDPs represent the network edges between many distinct pairs of brain regions, comprising in total 1,695 distinct region-pair brain connections (http://www.fmrib.ox.ac.uk/ukbiobank/). In addition to this being a very large number of IDPs from which to interpret association results, these individual IDPs tend to be substantially noisier than most of the other, more structural, IDPs. Hence, while we did carry out GWAS for each of these 1,695 connectivity IDPs, we also reduced the full set of connectivity IDPs into just six new summary IDPs using data-driven feature identification. We performed this dimensionality reduction by applying ICA[41], applied to all functional connectivity IDPs from all subjects, to find linear combinations of IDPs that are independent between the different features (ICA components) identified[42]. We carried out the ICA feature estimation without any use of the genetic data, and we maximized independence between component IDP weights (as opposed to subject weights). We used split-half reproducibility (across subjects) to optimize both the initial dimensionality reduction (14 eigenvectors from a singular value decomposition was found to be optimal) and also the final number of ICA components (6 ICA components was optimal, with reproducibility of ICA weight vectors greater than $r=0.9$). The resulting six ICA features were then treated as new IDPs, representing six independent sets (or, more accurately, linear combinations) of the original functional connectivity IDPs. These six new IDPs were added into the GWAS analyses. The six ICA features explain 4.9% of the total variance in the full set of network connection features, and are visualized in Supplementary Fig. 18. More details of the ICA analysis of the resting state data, together with browsing functionality of the highlighted brain regions can be found on the FMRIB UK Biobank Resource web page (http://www.fmrib.ox.ac.uk/ukbiobank/).

We organized all 3,144 IDPs into 9 groups (Supplementary Table 12), each with a distinct pattern of missing values (not all subjects have usable, high-quality data from all modalities[4]). For the GWAS in this study we did not try to impute missing IDPs owing to the low levels of correlation observed across groups.

The distributions of IDP values varied considerably between phenotype classes, with some phenotypes exhibiting substantial skew (Supplementary Fig. 19) that would probably invalidate the assumptions of the linear regression used to test for association. To ameliorate this, we quantile-normalized each of the IDPs before association testing. This transformation also helped to avoid undue influence of outlier values. We also (separately) tested an alternative process in which an outlier removal process was applied to the untransformed IDPs; this gave very similar results for almost all association tests, but was found to reduce the significance of a very small number of associations. This possible alternative method for IDP preprocessing was therefore not followed through (data not shown).

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Genetic data processing. We used the imputed genetic dataset made available by UK Biobank in its July 2017 release[6]. This consists of >92 million autosomal variants imputed from the Haplotype Reference Consortium (HRC) reference panel[43] and a merged UK10K + 1000 Genomes reference panel. We first identified a set of 12,623 participants who had also been imaged by UK Biobank. We then applied filters to remove variants with minor allele frequency (MAF) below 0.1% and with an imputation information score below 0.3, which reduced the number of SNPs to 18,174,817. We then kept only those samples (subjects) estimated to have recent British ancestry using the sample quality control information provided centrally by UK Biobank[6] (using the variable *in.white.British.ancestry.subset* in the file ukb_sqc_v2.txt); population structure can be a serious confound to genetic association studies[44], and this type of sample filtering is standard. This reduced the number of samples to 8,522. The UK Biobank dataset contains a number of close relatives (third cousins or closer). We therefore created a subset of 8,428 nominally unrelated subjects following procedures similar to those described previously[6]. After running GWAS on all the (SNP) variants in the 8,428 samples we applied three further variant filters to remove variants with a Hardy–Weinberg equilibrium P value $<10^{-7}$, remove variants with MAF <0.1% and keep only those variants in the HRC reference panel. This resulted in a dataset with 11,734,353 SNPs.

We used two separate datasets to replicate the associated variants found in this study. The first set of 930 subjects was a subset of the 1,279 subjects with imaging data that we did not use for the main GWAS, who had primarily been excluded because they were not in the recent British ancestry subset. An examination of these samples according the genetic principal components (PCs) revealed that many of those samples are mostly of European ancestry (Supplementary Fig. 20). We selected 930 samples with a first genetic PC <14 from Supplementary Fig. 20 and these constituted the replication sample. In January 2018 a further tranche of 4,588 samples with imaging data was released by UK Biobank. Of these subjects, we selected 3,956 subjects that both had genetic data available and also had been imaged in the same imaging centre as the discovery sample. We applied the same pre-processing pipeline as for the discovery set. We then restricted this to 3,456 subjects that were of recent British ancestry and replication tests were then conducted on these 3,456 subjects.

Potential confounds for brain IDP GWAS. There are a number of potential confounding variables when carrying out GWASs of brain IDPs. We used three sets of covariates in our analyses relating to (a) imaging confounds (b) measures of genetic ancestry, and (c) non-brain imaging body measures.

We identified a set of variables that were likely to represent imaging confounds, for example those associated with biases in noise or signal level, corruption of data by head motion or overall head size changes. For many of these we generated various versions (for example, using quantile normalization and also outlier removal, to generate two versions of a given variable, as well as including the squares of these to help model nonlinear effects of the potential confounds). This was done in order to generate a rich set of covariates and hence reduce as much as possible potential confounding effects on analyses such as the GWAS, which are particularly of concern when the subject numbers are so high[4,45].

Age and sex are can be variables of biological interest, but can also be sources of imaging confounds, and here were included in the confound regressors. Head motion is summarized from resting and task-based fMRI as the mean displacement (in mm) between one time point and the next, averaged over all time points and across the brain. Head motion can be a confounding factor for all modalities and not just those comprising timeseries of volumes, but is readily estimable only from the timeseries modalities. Nevertheless, the amount of head motion is expected to be reasonably similar across all modalities (for example, correlation between head motion in resting and task fMRI is $r=0.52$) and so it is worth using fMRI-derived head motion estimates as confound regressors for all modalities.

The exact location of the head and the radio-frequency receiver coil in the scanner can affect data quality and IDPs. To help to account for variations in position in different scanned participants, several variables have been generated that describe aspects of the positioning (see http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25756, http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25757, http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25758, and http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=25759). The intention is that these can be useful as 'confound variables'; for example, these might be regressed out of brain IDPs before carrying out correlations between IDPs and non-imaging variables. TablePosition is the Z-position of the coil (and the scanner table on which the coil sits) within the scanner (the Z axis points down the centre of the magnet). BrainCoGZ is somewhat similar, being the Z-position of the centre of the brain within the scanner (derived from the brain mask estimated from the T1-weighted structural image). BrainCoGX is the X-position (left–right) of the centre of the brain mask within the scanner. BrainBackY is the Y-position (front–back relative to the head) of the back of brain mask within the scanner.

UK Biobank brain imaging aims to maintain as fixed an acquisition protocol as possible during the 5–6 years that the scanning of 100,000 participants will take. There have been a number of minor software upgrades (the imaging study seeks to minimize any major hardware or software changes). Detailed descriptions of every protocol change, along with thorough investigations of the effects of these on the resulting data, will be the subject of a future paper. Here, we attempted to model

any long-term (over scan date) changes or drifts in the imaging protocol or software or hardware performance, by generating a number of data-driven confounds. The first step was to form a temporary working version of the full subjects × IDPs matrix with outliers limited (see below) and no missing data, using a variant of low-rank matrix imputation with soft thresholding on the eigenvalues[46]. Next, the data were temporally regularized (approximate scale factor of several months with respect to scan date, see https://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=53, Instance 2) with spline-based smoothing. We then applied PCA and kept the top 10 components, to generate a basis set that reflects the primary modes of slowly changing drifts in the data.

To describe the full set of imaging confounds we use a notation where subscript i indicates quantile normalization of variables, and m indicates median-based outlier removal (discarding values greater than five times the median absolute deviation from the overall median). If no subscript is included, no normalization or outlier removal was carried out. Certain combinations of normalization and powers were not included, either because of very high redundancy with existing combinations, or because a particular combination was not well-behaved. The full set of variables used to create the confounds matrix are: a, age at time of scanning, demeaned (cross-subject mean subtracted); s, sex, demeaned; q, four confounds relating to the position of the radio-frequency coil and the head in the scanner (see above), all demeaned; d, ten drift confounds (see above); m, two measures of head motion (one from resting fMRI, one from task-based fMRI); and h, volumetric scaling factor needed to normalize for head size[47].

The full matrix of imaging confounds is then:

$$[a \;\; a^2 \;\; a\times s \;\; a^2\times s \;\; a_i \;\; a_i^2 \;\; a_i\times s \;\; a_i^2\times s \;\; m_m \;\; m_m^2 \;\; h_m \;\; q_m \;\; q_m^2 \;\; d_m \;\; m_i \;\; h_i \;\; q_i \;\; q_i^2 \;\; d_i]$$

Any missing values in this matrix are set to zero after all columns have had their mean subtracted. This results in a full-rank matrix of 53 columns (ratio of maximum to minimum eigenvalues is 42.6). Additional discussion on the dangers and interpretation of imaging confounds in big imaging data studies, particularly in the context of disease studies, has been published[45].

Genetic ancestry is a well-known potential confound in GWAS. We ameliorated this by filtering out samples that were not of recent British ancestry. However, a set of 40 genetic principal components (PCs) has been provided by UK Biobank[6], and we used these PCs as covariates in all of our analyses. The matrix of imaging confounds, together with a matrix of 40 genetic principal components, was regressed out of each IDP before the analyses reported here.

There exist a number of substantial correlations between IDPs and non-genetic variables collected on the UK Biobank subjects[4]. We therefore also carried out some analyses involving variables relating to blood pressure (diastolic and systolic), height, weight, head bone mineral density, head bone mineral content and two principal components from the broader set of bone mineral variables available (https://biobank.ctsu.ox.ac.uk/crystal/docs/DXA_explan_doc.pdf). Supplementary Fig. 21 shows the association of these eight variables against the IDPs and shows significant associations. These are variables that are likely to have a genetic basis, at least in part. Genetic variants associated with these variables might then produce false positive associations for IDPs. To investigate this possibility, we ran GWASs for these eight traits (conditioned on the imaging confounds and genetic PCs) (Supplementary Fig. 22). We also ran a parallel set of IDP GWASs with these 'body confounds' regressed out of the IDPs.

Heritability and genetic correlation of IDPs. We used a linear mixed model implemented in the SBAT (sparse Bayesian association test) software (https://jmarchini.org/sbat/) to calculate additive genetic heritabilities for the $P=3{,}144$ traits. To estimate genetic correlations we used a multi-trait mixed model. If Y is an $N\times P$ matrix of P phenotypes (columns) measured on N individuals (rows) then we use the model:

$$Y = U + \varepsilon \tag{1}$$

where U is an $N\times P$ matrix of random effects and ε is an $N\times P$ matrix of residuals, and these are modelled using Matrix normal distributions as follows:

$$U \sim MN(0, K, B)$$

$$\varepsilon \sim MN(0, I_N, E)$$

In this model, K is the $N\times N$ kinship matrix between individuals, B is the $P\times P$ matrix of genetic covariances between phenotypes and E is the $P\times P$ matrix of residual covariances between phenotypes. We estimate the covariance matrices B and E using a new C++ implementation of an EM algorithm[48] included in the SBAT software (https://jmarchini.org/sbat/).

For the marginal heritabilities and genetic correlation analysis we used a realized relationship matrix (RRM) for the kinship matrix (K). This RRM was calculated from the 8,428 nominally unrelated individuals using fastLMM (https://github.com/MicrosoftGenomics/FaST-LMM). We used the subset of imputed SNPs that were both assayed by the genotyping chips and included in the HRC reference panel, and so will essentially be hard-called genotypes. In addition, all SNPs with duplicate rsids (reference SNP cluster IDs) were removed. PLINK (http://www.cog-genomics.org/plink/2.0/) was used for file conversion before input into fastLMM.

To estimate genetic correlations, we fit the model to several of the groupings of IDPs detailed in Supplementary Table 12. The estimated covariance matrices B and E were used to estimate the genetic correlation of pairs of IDPs. The genetic correlation between the ith and jth IDPs in a jointly analysed group of IDPs is estimated as

$$r_{ij} = \frac{B_{ij}}{\sqrt{B_{ii}B_{jj}}}$$

Multi-trait association tests. We used a multi-trait mixed model to test each SNP for association with different groupings of traits (Supplementary Table 7). The model has the form $Y = G\alpha + U + \varepsilon$, where G is an $N\times 1$ vector of SNP dosages and α is a $1\times P$ vector of effect sizes. We fit the model using estimates of B and E from the 'null' model with $\alpha = 0$ and a leave one chromosome out (LOCO) approach for RRM calculation. We ran this test on the main set of 8,428 samples and on the replication samples. For the replication analysis we used the estimates of B and E from the main set of 8,428 samples. This test was implemented in SBAT software.

Genetic association of IDPs. We used BGENIE v1.2 (https://jmarchini.org/bgenie/) to carry out GWASs of imputed variants against each of the processed IDPs. This program was designed to carry out the large number of IDP GWAS required in this analysis. It avoids repeated reading of the genetic data file for each IDP and uses efficient linear algebra libraries and threading to achieve good performance. The program has already been used by several studies to analyse genetic data from the UK Biobank[49,50]. We fit an additive model of association at each variant, using expected genotype count (dosage) from the imputed genetic data. We ran associated tests on the main set of 8,428 samples and the replication samples.

Identifying associated genetic loci. Most GWAS analyse only one or a few different phenotypes, and often uncover just a handful of associated genetic loci, which can be interrogated in detail. Owing to the large number of associations uncovered in this study, we developed an automated method to identify, distinguish and count individual associated loci from the 3,144 GWASs (one GWAS for each IDP). For each GWAS we first identified all variants with $-\log_{10}(P) > 7.5$. We applied an iterative process that starts by identifying the most strongly associated variant, storing it as a lead variant, and then removing it, and all variants within 0.25 cM from the list of variants (equivalent to approximately 250 kb in physical distance). The process was then repeated until the list of variants was empty. We applied this process to each GWAS using two filters on MAF: (a) MAF > 0.1%, and (b) MAF > 1%. We grouped associated lead SNPs across phenotypes into clusters. This process first grouped SNPs within 0.25 cM of each other, and this mostly produced sensible clusters, but some hand curation was used to merge or split clusters based on visual inspection of cluster plots and levels of linkage disequilibrium between SNPs. For some clusters in Extended Data Table 1, we report coding SNPs that were found to be in high linkage disequilibrium with the lead SNPs.

Accounting for multiple IDPs. We adjusted the genome-wide significance threshold ($-\log_{10}(P) > 7.5$) by a Bonferroni factor ($-\log_{10}(3{,}144) = 3.5$) that accounts for the number of IDPs tested, giving a threshold of $-\log_{10}(P) > 11$. This assumes (incorrectly) that the IDPs are independent and so is likely to be conservative, but we preferred to be cautious when analysing so many IDPs.

Genetic correlation analysis. We used linkage disequilibrium score regression[51] to estimate the genetic correlation between the IDPs studied in our analysis and ten disease-, personality- or brain-related traits. We gathered summary statistics for GWASs of the neuroticism personality trait (https://www.thessgac.org/data), autism spectrum (https://www.med.unc.edu/pgc/) and sleep duration (http://www.t2diabetesgenes.org/data/) and also seven disease traits: attention deficit hyperactivity disorder, schizophrenia, major depressive disorder and bipolar disorder (https://www.med.unc.edu/pgc/), Alzheimer's disease (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php), stroke (PMC4818561 from http://cerebrovascularportal.org/informational/downloads) and amyotrophic lateral sclerosis (http://databrowser.projectmine.com/). The number of samples in each of these studies and the DOIs for the corresponding studies are provided in Supplementary Table 13.

For each IDP–trait pair, we used the LDSCORE regression software (v1.0.0; https://github.com/bulik/ldsc) to compute the genetic correlation between the IDP and the trait, with linkage disequilibrium measurements taken from the 1000 Genomes Project (provided by the maintainers of the LDSCORE regression

software). We filtered the SNPs to include only those with imputation INFO ≥ 0.9 and MAF $\geq 0.1\%$. Only INFO scores for major depressive disorder, schizophrenia and attention deficit hyperactivity disorder were provided by the source studies, and so for these three analyses we applied the INFO threshold to both the SNPs from our study and also the source study. For the remaining six studies, an INFO filter was applied to the SNPs from our own study. Owing to low levels of heritability of the functional edge IDPs, all of these were removed from this analysis. As calculation of genetic correlation between traits only really makes sense if both traits are themselves heritable, we only used those IDPs with z-scores for significantly non-zero heritability greater than 4. In total, we used 897 IDPs. To account for correlations between IDPs, we used the raw phenotype correlation matrix to simulate z-scores (and associated tail probabilities) using samples from a multivariate normal distribution with that same correlation matrix.

Analysis of enrichment of functional categories. We used the LDSCORE regression software to carry out the heritability enrichment partitioning analysis into different functional categories (https://github.com/bulik/ldsc). We used 24 functional categories: coding, UTR, promoter, intron, histone marks H3K4me1, H3K4me3, H3K9ac5 and two versions of H3K27ac, open chromatin DNase I hypersensitivity site (DHS) regions, combined chromHMM/Segway predictions, regions conserved in mammals, super-enhancers and active enhancers from the FANTOM5 panel of samples. For each IDP, the enrichment of each functional category was summarized as the proportion of h^2 explained by the category divided by the proportion of common variants in the category. For each IDP and each annotation we used the two-sided enrichment P value as reported by the LDSCORE regression software. We labelled those P values as enriched or depleted depending on whether the enrichment estimate was greater or less than 1. We stratified these P values accordingly into 23 groups of IDPs.

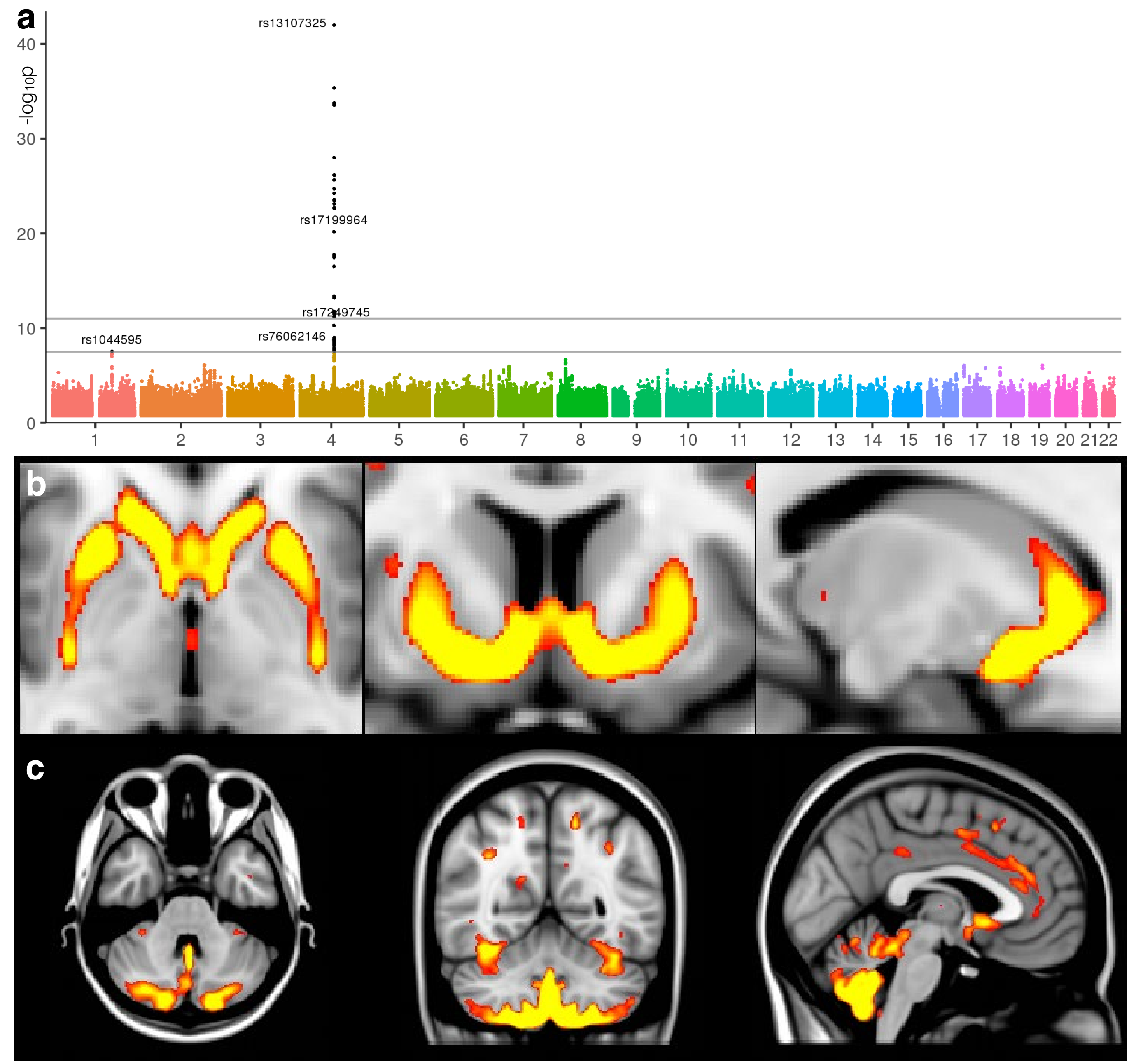
Code availability. Most of the software and code used in this study are publicly available, including custom Matlab scripts used to prepare IDPs for GWAS (http://www.fmrib.ox.ac.uk/ukbiobank/gwaspaper/). Pre-compiled binaries for the latest version of BGENIE and SBAT are available at https://jmarchini.org/software/. This software is currently licensed free for use by researchers at academic institutions. Commercial organizations wishing to use these packages must enquire about a licence from the University of Oxford. Brain image processing was largely carried out with FSL (FMRIB's Software Library, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki) and further Matlab-based preparation of IDPs and imaging confounds utilized code from FSLNets (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLNets).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

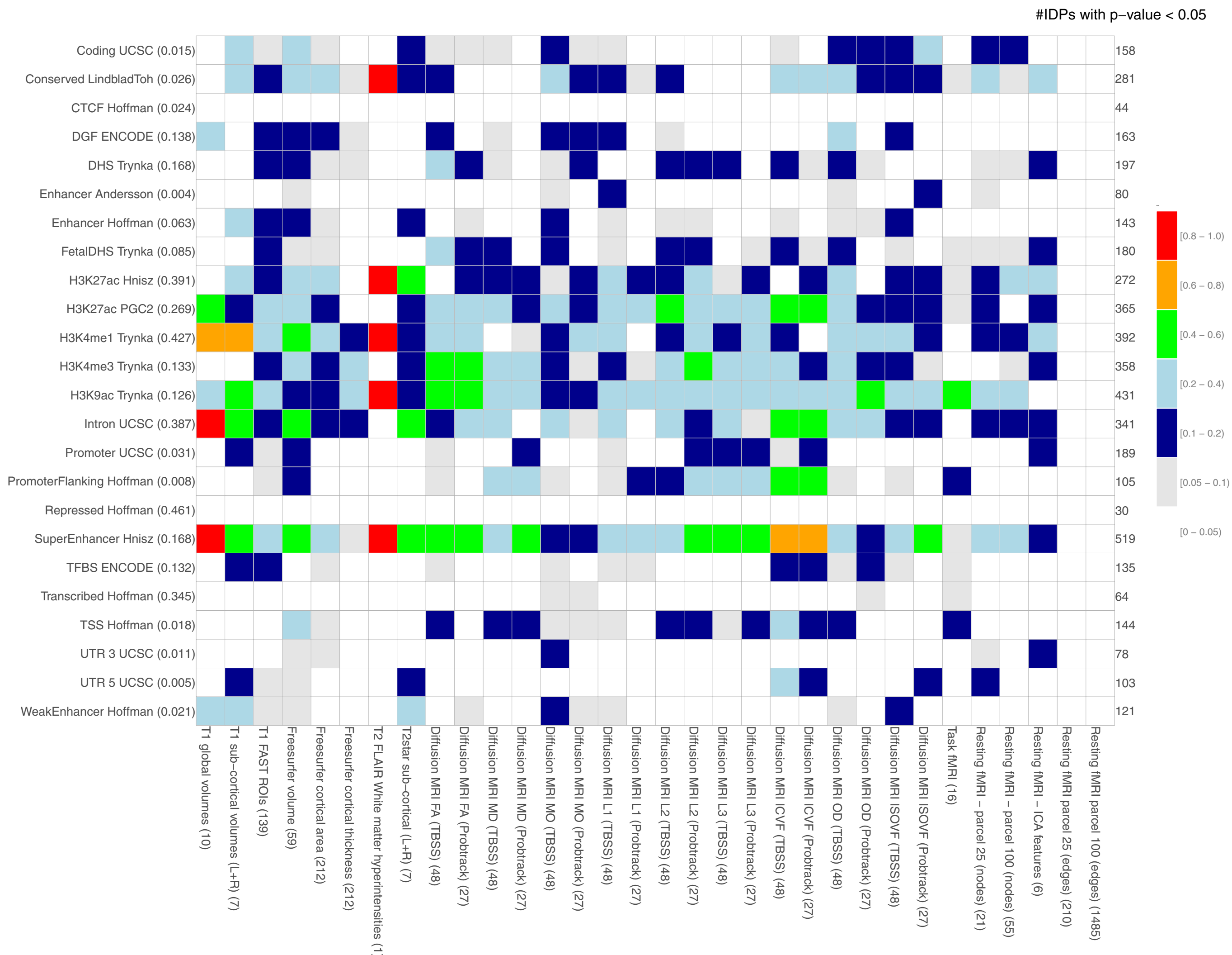
Data availability

The full set of GWAS results from this study is available on the Oxford BIG web browser (http://big.stats.ox.ac.uk/), which allows users to browse associations by SNP, gene or phenotype.

37. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).
38. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9**, 195–207 (1999).
39. Klein, A. & Tourville, J. 101 labeled brain images and a consistent human cortical labeling protocol. *Front. Neurosci.* **6**, 171 (2012).
40. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53**, 1–15 (2010).
41. Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10**, 626–634 (1999).
42. Duff, E. P. et al. Learning to identify CNS drug action and efficacy using multistudy fMRI data. *Sci. Transl. Med.* **7**, 274ra16 (2015).
43. The Haplotype Reference Consortium A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Med.* **48**, 1279–1283 (2015).
44. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
45. Smith, S. M. & Nichols, T. E. Statistical challenges in "big data" human neuroimaging. *Neuron* **97**, 263–268 (2018).
46. Cai, J.-F., Candès, E. J. & Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20**, 1956–1982 (2010).
47. Smith, S. M. et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* **17**, 479–489 (2002).
48. Dahl, A. et al. A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* **48**, 466–472 (2016).
49. Luciano, M. et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50**, 6–11 (2017).
50. Davies, G. et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun.* **9**, 2098 (2018).
51. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).



Extended Data Fig. 1 | Manhattan plot and spatial mapping of the associations between grey matter volume and rs13107325 (*SLC39A8*). **a**, The Manhattan plot relates to the original GWAS for the IDP of grey matter volume in the left ventral striatum. **b**, **c**, Spatial mapping of rs13107325 against voxelwise local grey matter volume (grey matter was averaged across all 1,181 subjects with one copy of the non-reference allele, and the average from all 7,215 subjects that had zero copies was subtracted from that, for display in colour here; the difference was thresholded at 0.015 (unitless relative measure of local grey matter volume)). The maps show that the effect of rs13107325 is found more generally bilaterally in the ventral caudate, putamen, ventral striatum, anterior cingulate cortex, and with a strong cerebellar contribution (lobules VI–X), particularly in the prefrontal-projecting Crus I/II, which are selectively expanded in humans.



Extended Data Fig. 2 | Partitioning of heritability by functional category. The plot shows the proportion of IDPs in each of the 23 IDP groupings (*x*-axis) that show a nominal enrichment *P* value <0.05 (two-sided tests, uncorrected *P* values, see Methods) for the 24 functional categories (*y*-axis). The total number of such IDPs for each category is given on the right edge of the plot. The number of IDPs in each IDP group is given in parentheses in the *x*-axis labels. The proportion of the genome annotated by each functional category is given in parentheses in the *y*-axis labels.

Extended Data Table 1 | Summary of most highly associated SNP–IDP clusters

cluster index	cluster name	# IDPs	top IDP	chr	RSID	position	locus	ref allele	nonref allele	nonref AF	p value	replication p-value (N=3456)	replication p-value (N=930)	GTEX eQTL
1	Volume Cerebellum VIIIa (vermis)	1	T1_FAST_ROIs_V_cerebellum_VIIIa	1	rs76934732	76013268	SLC44A5	G	A	0.145	8.51E-13	6.10E-04	5.22E-02	SLC44A5 ACADM
2	dMRI Corpus callosum (genu)	1	dMRI_TBSS_ICVF_Genu_of_corpus_callosum	1	rs2365715	156615114	BCAN	A	G	0.388	5.38E-12	4.50E-03	1.33E-02	BCAN. APOA1BP, SYT11
3	Volume WM lesions	1	T2_FLAIR_BIANCA_WMH_volume	2	rs3762515 (5' UTR)	56150864	EFEMP1	C	T	0.0959	4.27E-13	1.18E-02	4.84E-01	
4	rfMRI Cortical and cerebellar motor nodes and edges	2	NODEamps25_0012	2	rs60873293	114092549	intergenic	G	T	0.217	9.86E-15	3.10E-07	9.50E-02	AC016745.3, RP11-480C16.1
5	T2* Pallidum	1	SWI_T2*_pallidum_L+R	2	rs6740926	190326498	WDR75	C	T	0.038	1.31E-14	3.50E-09	3.78E-04	WDR75
6	rfMRI Middle temporal sulcus nodes and edges	2	netmat_ICA_003	3	rs35124509 (missense)	89521693	EPHA3	T	C	0.3853	4.49E-22	3.27E-09	3.73E-03	EPHA3
7	T2* Putamen and pallidum	6	SWI_T2*_putamen_L+R	3	rs4428180	133466374	TF	A	G	0.152	2.23E-22	6.11E-07	1.03E-03	TF
8	rfMRI Prefrontal and parietal edges	1	netmat_ICA_002	3	rs2279829 (3' UTR)	147106319	ZIC4	C	T	0.221	8.34E-12	5.46E-05	2.51E-03	
9	dMRI Superior cerebellar peduncles	8	dMRI_TBSS_ICVF_Superior_cerebellar_peduncle_L	4	rs4697414	23724255	RP11-380P13.2	C	T	0.823	5.83E-24	1.33E-06	4.63E-02	RP13-497K6.1, RP11-380P13.2
10	Volume Putamen, ventral striatum, cerebellum VIIIb, IX, X; T2* Pallidum; dMRI Cerebral peduncles	20	IDP_T1_FAST_ROIs_L_ventral_striatum	4	rs13107325 (missense)	103188709	SLC39A8	C	T	0.073	1.04E-42	6.64E-20	8.97E-06	
11	dMRI Most WM tracts	199	dMRI_ProbtrackX_ICVF_ilf_r	5	rs67827860	82860485	VCAN	C	T	0.188	4.06E-37	3.93E-12	2.19E-04	
12	rfMRI Parietal and prefrontal edges	1	netmat_ICA_004	5	rs7442779	92788278	NR2F1-AS1	A	G	0.05	8.18E-15	1.90E-04	4.04E-02	
13	dMRI Corpus callosum (genu, body, splenium)	7	dMRI_TBSS_ICVF_Genu_of_corpus_callosum	5	rs4150221	139719991	HBEGF	T	C	0.264	8.43E-20	1.72E-09	4.06E-02	SRA1
14	T2* Putamen	3	SWI_T2*_putamen_L+R	6	rs1800562 (missense)	26093141	HFE	G	A	0.0768	6.61E-20	2.91E-04	3.44E-03	U91328.19
15	dMRI Crossing pontine tract	1	dMRI_TBSS_MO_Pontine_crossing_tract	7	rs2286184	84630516	SEMA3D	C	T	0.201	5.31E-17	6.02E-09	1.58E-04	
16	dMRI Corpus callosum (genu)	1	dMRI_TBSS_OD_Genu_of_corpus_callosum	7	rs12113919	117612315	intergenic	C	G	0.416	3.96E-12	1.44E-04	1.84E-03	CTTNBP2
17	Volume Brain	2	volume_MaskVol	7	rs2908004 (missense)	120969769	WNT16	G	A	0.4455	3.55E-16	7.07E-09	2.50E-04	CPED1, FAM3C
18	T2* Putamen	2	SWI_T2*_putamen_L+R	8	rs35469695	23406169	SLC25A37	C	G	0.174	2.22E-12	2.11E-02	2.17E-01	SLC25A37
19	Volume Pallidum	3	T1_FIRST_pallidum_volume_L+R	8	rs2923405	42448126	SMIM19/SLC20A2	T	G	0.583	3.31E-17	1.34E-04	5.98E-03	SMIM19, SLC20A2
20	T2* Pallidum	2	SWI_T2*_pallidum_L+R	8	rs2978098	101676675	SNX31	A	C	0.468	6.43E-15	1.08E-05	3.23E-01	SNX31
21	Volume Cerebellum	3	T1_FAST_ROIs_L_cerebellum_crus_I	9	rs72754248	119061396	PAPPA	G	A	0.069	1.38E-17	4.23E-06	2.01E-01	
22	T2* Pallidum, putamen and caudate	17	SWI_T2*_pallicum_L+R	10	rs10764176 (missense)	18,242,311	SLC39A12	A	G	0.3	3.30E-21	1.01E-11	9.71E-02	SLC39A12
23	T2* Caudate	3	SWI_T2*_caudate_L+R	10	rs12570727	18,425,519	CACNB2	G	A	0.394	2.17E-22	2.20E-10	6.23E-04	SLC39A12-AS1
24	rfMRI Parietal, temporal and prefrontal nodes	20	NODEamps100_0002	10	rs2274224 (missense)	96039597	PLCE1	G	C	0.431	6.55E-19	1.73E-03	7.21E-02	NOC3L, PLCE1, PLCE1-AS1
25	rfMRI Prefrontal nodes	6	NODEamps25_0013	10	rs11596664	134280157	INPP5A	C	T	0.439	1.97E-15	2.23E-05	3.60E-02	INPP5A RP11, 432J24.6
26	T2* Pallidum	3	SWI_T2*_pallidum_L+R	11	rs11230859	61769972	intergenic	G	A	0.663	2.31E-17	6.39E-03	4.83E-02	
27	dMRI Crossing pontine tract	1	dMRI_TBSS_MO_Pontine_crossing_tract	11	rs4935898 (missense)	124742385	ROBO3	G	A	0.048	1.76E-19	2.47E-05	2.47E-01	
28	Volume Mesencephalon (WM cerebellum, brainstem)	3	volume_Right-Cerebellum-White-Matter	12	rs4301837	102336310	DRAM1 GNPTAB CHPT1	T	C	0.501	3.40E-13	3.37E-04	1.23E-02	GNPTAB, CHPT1, DRAM1
29	Volume Hippocampus	2	T1_FAST_ROIs_R_hippocampus	12	rs7315280	117320938	intergenic	A	G	0.115	7.06E-14	6.80E-05	6.69E-01	FBXW8, HRK
30	Volume Putamen	4	volume_Right-Putamen	14	rs945270	56200473	intergenic	C	G	0.419	3.67E-14	9.27E-06	3.32E-03	
31	Volume and area of precuneus and cuneus	11	T1_FAST_ROIs_R_intracalc_cortex	14	rs74826997	59628609	DAAM1	T	C	0.125	2.46E-16	3.08E-07	2.88E-02	L3HYPDH, JKAMP
32	Thickness, area and volume of primary sensorimotor cortex	15	a2009s_lh_S_central_area	15	rs4924345	39639898	RP11-624L4.1	A	C	0.081	3.27E-53	1.69E-27	1.01E-06	
33	Volume 4th ventricle	1	volume_4th-Ventricle	15	rs2642636	58363242	ALDH1A2	C	G	0.415	5.24E-16	5.63E-03	1.81E-01	ALDH1A2, AQP9
34	dMRI Uncinate	4	dMRI_ProbtrackX_ISOVF_unc_r	16	rs7197215	51449978	intergenic	A	G	0.566	2.24E-15	4.50E-02	1.43E-04	
35	Volume Cerebellum IX	2	T1_FAST_ROIs_L_cerebellum_IX	17	rs9905515	35261073	RP11-445F12.1	G	C	0.23	3.32E-13	9.84E-06	2.70E-04	
36	T2* Caudate and putamen	6	SWI_T2*_putamen_L+R	17	rs668799	40716235	COASY	C	T	0.278	1.43E-17	1.79E-04	9.86E-04	TUBG2, CNTNAP1, FAM134C, NAGLU, BECN1, HSD17B1, PLEKHH3
37	Volume WM lesions	1	T2_FLAIR_BIANCA_WMH_volume	17	rs3744020	73871773	TRIM47	G	A	0.188	1.15E-12	6.05E-06	3.36E-02	TRIM47, TRIM65, RP11-552F3.9, etc.
38	dMRI Crossing pontine tract	1	dMRI_TBSS_MO_Pontine_crossing_tract	18	rs2928990	49421125	intergenic	T	G	0.898	3.97E-16	3.96E-05	2.27E-03	

The table summarizes the 38 clusters of SNP–IDP associations ($n = 8{,}428$ subjects, see main text and Methods for details). For each cluster, the most significant association between an SNP and an IDP is detailed by the chromosome, rsID, base-pair position, SNP alleles, non-reference allele frequency, *P* value in the discovery sample and the replication *P* values. The locus column details a gene if the SNP is in that gene. If we found a coding SNP or eQTL in high linkage disequilibrium with the lead SNP, then this is reported instead.

https://doi.org/10.1038/s41586-018-0461-z

Accurate classification of *BRCA1* variants with saturation genome editing

Gregory M. Findlay[1], Riza M. Daza[1], Beth Martin[1], Melissa D. Zhang[1], Anh P. Leith[1], Molly Gasperini[1], Joseph D. Janizek[1], Xingfan Huang[1], Lea M. Starita[1,2]* & Jay Shendure[1,2,3]*

Variants of uncertain significance fundamentally limit the clinical utility of genetic information. The challenge they pose is epitomized by *BRCA1*, a tumour suppressor gene in which germline loss-of-function variants predispose women to breast and ovarian cancer. Although *BRCA1* has been sequenced in millions of women, the risk associated with most newly observed variants cannot be definitively assigned. Here we use saturation genome editing to assay 96.5% of all possible single-nucleotide variants (SNVs) in 13 exons that encode functionally critical domains of BRCA1. Functional effects for nearly 4,000 SNVs are bimodally distributed and almost perfectly concordant with established assessments of pathogenicity. Over 400 non-functional missense SNVs are identified, as well as around 300 SNVs that disrupt expression. We predict that these results will be immediately useful for the clinical interpretation of *BRCA1* variants, and that this approach can be extended to overcome the challenge of variants of uncertain significance in additional clinically actionable genes.

Our ability to predict the phenotypic consequences of an arbitrary genetic variant in a human genome remains poor. This problem is evidenced by the large numbers of variants of uncertain significance (VUS) identified in 'actionable' genes, that is, genes in which the definitive identification of a pathogenic variant would alter clinical management[1]. For example, heterozygous germline variants that disrupt *BRCA1* markedly increase the risk of early-onset breast and ovarian cancer[2,3] and are actionable, as more frequent screening or prophylactic surgery can lead to improved outcomes[4,5]. Clinical sequencing can identify specific variants as risk-conferring[6]. However, as of January 2018, most *BRCA1* SNVs are classified as VUS[7]. VUS are typified by rare missense SNVs, but also include variants potentially affecting messenger RNA (mRNA) levels. Further illustrating the challenge associated with VUS, there are hundreds of *BRCA1* SNVs that have received conflicting interpretations[7].

There are two main approaches for resolving VUS. The first approach, data sharing, relies on the expectation that as *BRCA1* is sequenced in more individuals, the recurrent observation of a variant in individuals who either have or have not developed cancer will enable its interpretation. However, given that the majority of potential variants in *BRCA1* are extremely rare and that the phenotype is incompletely penetrant, it is unclear whether sufficient numbers of humans will ever be sequenced to accurately quantify cancer risk for each possible variant.

The second approach, functional assessment, has spurred the development of diverse in vitro assays for *BRCA1*[8]. As the homology-directed DNA repair (HDR) function of BRCA1 is key for tumour suppression, one commonly used assay measures whether expression of a BRCA1 variant can rescue HDR integrity[9,10]. Other BRCA1 assays evaluate embryonic stem cell viability[11], transcriptional activation[12], drug sensitivity[11], protein–protein interaction[9,13] or splicing[14,15]. Computational predictions based on features such as conservation can be informative but are insufficiently accurate to be used in the absence of genetic or experimental evidence[16].

Experimental assessments of *BRCA1* variants have been limited in several ways. First, they are typically performed post hoc and have not kept pace with the discovery of VUS. Second, assays expressing variants as cDNA-based transgenes removed from their genomic context[9,13] fail to assess the effects on splicing or transcript stability, and risk artefacts of overexpression[17]. Genome editing provides a potential means to overcome these challenges, but has yet to be applied to characterize any appreciable number of VUS in *BRCA1* or other genes similarly linked to cancer predisposition.

Here we set out to apply genome editing to measure the functional consequences of all possible SNVs in key regions of *BRCA1*, regardless of whether they have been previously observed in a human. Given the large size of *BRCA1*, we prioritized 13 exons that encode the RING and BRCT domains, which critically underlie its role as a tumour suppressor[18–20]. In addition to around 400 VUS or variants with conflicting interpretations, all 21 *BRCA1* missense SNVs classified by a ClinVar-approved expert panel as pathogenic reside in these exons[7], as do missense and splice variants shown to disrupt *BRCA1* in functional assays[11,21] (ClinVar is a widely used database of clinical variant interpretations submitted by clinical testing laboratories). In each experiment, a single exon is subjected to saturation genome editing (SGE)[22], wherein all possible SNVs are simultaneously introduced and concurrently assayed. We used SGE to measure functional effects for 3,893 SNVs, comprising 96.5% of all possible SNVs in the targeted exons. These scores are bimodally distributed and nearly perfectly concordant with expert-based assessments of pathogenicity. We predict that our functional classifications will be of immediate clinical utility, and that scaling this approach to additional genes will substantially enhance the utility of genetic testing.

Saturation genome editing of *BRCA1* exons

Many genes in the HDR pathway, including the hereditary cancer predisposition genes *BRCA1*, *BRCA2*, *PALB2* and *BARD1*[6], have been deemed essential in the human haploid cell line HAP1[23] (Fig. 1a). To confirm this, we transfected HAP1 cells with a plasmid co-expressing Cas9 and guide RNAs (gRNAs) targeting each of these genes[24]. High cell death was evident by light microscopy, and a luminescence-based survival assay established that targeting any of these genes substantially reduces HAP1 viability (Extended Data Fig. 1a–c). Deep sequencing of

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. *e-mail: lstarita@uw.edu; shendure@uw.edu

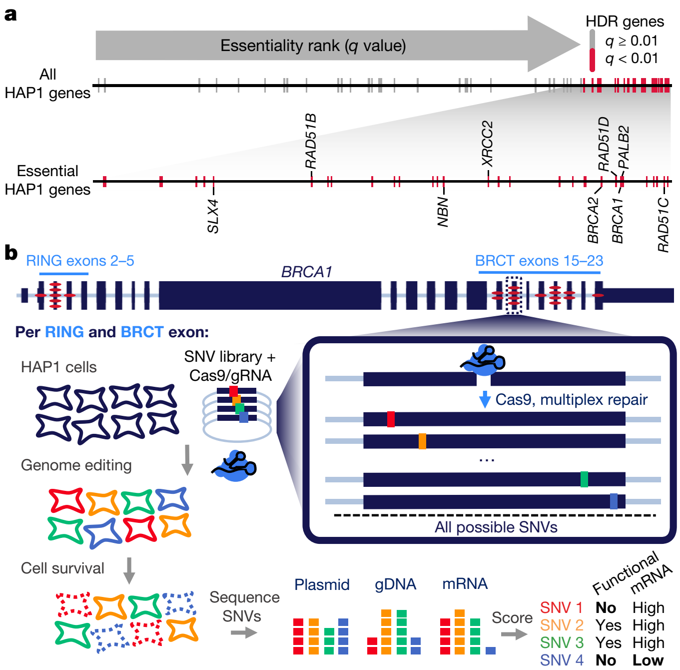


Fig. 1 | *BRCA1* and other HDR pathway genes are essential in HAP1 cells. a, The q-value rankings[23] of HDR pathway genes ($n = 66$) among 14,306 genes scored in a HAP1 gene trap screen for essentiality are indicated with tick marks. Essential HDR genes are coloured red and those implicated in cancer predisposition are labelled in the enlargement below. Of the 66 HDR pathway genes scored, 34 including *BRCA1* were 'essential', a 3.4-fold enrichment compared to non-HDR genes (Fisher's exact test, $P = 6.1 \times 10^{-12}$). **b**, SGE experiments were designed to introduce all possible SNVs across 13 *BRCA1* exons encoding the RING (exons 2–5, NCBI, NM_007294.3) and BRCT domains (exons 15–23). The exonic locations of all 21 *BRCA1* missense variants in ClinVar deemed pathogenic by an expert panel are indicated by red ovals. For each exon, a Cas9/gRNA construct was transfected with a library of plasmids containing all SNVs within approximately 100 bp of genomic sequence (the 'SNV library'). SNV library plasmids contained homology arms, as well as fixed synonymous variants within the CRISPR target site to prevent re-cutting. Upon transfection, successfully edited cells carried a single *BRCA1* SNV from the library. Cells were sampled 5 and 11 days after transfection and targeted gDNA and RNA sequencing was performed to quantify SNV abundances. SNVs compromising *BRCA1* function were selected against, manifesting in reduced gDNA representation, and SNVs that affect mRNA production were depleted in RNA relative to gDNA.

the edited loci of *BRCA1*-targeted cells confirmed that cell death was consequent to mutations, as there was widespread selection against frameshifting indels (Extended Data Fig. 1d). Overall, these results confirm the importance of HDR pathway components in HAP1 cells.

We next designed and optimized experiments for SGE[22] (Fig. 1b), focusing on the 13 exons of *BRCA1* that encode the RING and BRCT domains (exons 2–5 and 15–23, respectively; NCBI, NM_007294.3). To create libraries of repair templates, we used array-synthesized oligonucleotide pools containing all possible SNVs spanning each exon and around 10 base pairs (bp) of adjacent intronic sequence. Oligonucleotide pools for each exon were cloned into plasmids with homology arms ('SNV libraries'). Each design also included a fixed synonymous substitution at the Cas9 target site to reduce re-cutting after successful HDR[22]. Each SGE experiment targeted one exon. A population of 20 million HAP1 cells was co-transfected on day 0 with a corresponding SNV library and Cas9/gRNA plasmid. Variant frequencies were quantified by targeted sequencing of the edited exon from genomic DNA (gDNA) collected on day 5 and day 11.

We initially performed SGE in replicate for each exon in wild-type HAP1 cells. In each exon, we observed the expected depletion of frameshifting indels (Extended Data Fig. 2). However, to achieve more robust data, we optimized SGE in HAP1 cells in two ways. First, to increase HDR rates[25], we generated a monoclonal *LIG4*-knockout HAP1 line (HAP1-LIG4^{KO}) (Extended Data Fig. 3a, g). Second, as HAP1 cells can spontaneously revert to diploidy[26], sorting HAP1 cells for 1n ploidy before editing improved reproducibility (Extended Data Fig. 3b, h).

We performed optimized SGE on each of the 13 exons in 1n-sorted HAP1-LIG4^{KO} cells. We observed a median 3.6-fold increase in HDR rates on day 5 in HAP1-LIG4^{KO} relative to wild-type HAP1 cells (Fig. 2a), allowing us to test nearly every SNV in replicate (Extended Data Fig. 3c). Because these optimizations increased reproducibility without substantially altering SNV effects on survival (Fig. 2b, Extended Data Figs. 3, 4), we proceeded with data from the 1n-sorted HAP1-LIG4^{KO} cells. Additionally, targeted RNA sequencing of day 5 HAP1-LIG4^{KO} samples was used to determine the abundance of exonic SNVs in *BRCA1* mRNA (Fig. 2c).

Function scores for 3,893 *BRCA1* SNVs

To calculate function scores for each SNV, we first calculated the $\log_2$ ratio of the frequency of a SNV on day 11 to its frequency in the plasmid library. Second, positional biases in editing rates were modelled using day 5 SNV frequencies and subtracted (Extended Data Fig. 5). Third, to enable comparisons between exons, we normalized function scores such that the median synonymous and nonsense SNV in each experiment matched global medians. Lastly, a small number of SNVs that could not confidently be scored were filtered out (Extended Data Fig. 6). Altogether, we obtained function scores for 3,893 SNVs, which comprise 96.5% of all possible SNVs within or immediately intronic to these exons (Supplementary Table 1; https://sge.gs.washington.edu/BRCA1/).

Function scores were bimodally distributed (Fig. 2d). All nonsense SNVs scored below −1.25 ($n = 138$, median = −2.12), whereas 98.7% of synonymous SNVs more than 3 bp from splice junctions scored above −1.25 ($n = 544$, median = 0.00). We classified all SNVs as 'functional', 'non-functional', or 'intermediate' by fitting a two-component Gaussian mixture model (Extended Data Fig. 7). We categorized 72.5% of SNVs as functional, 21.1% as non-functional and 6.4% as intermediate.

It is particularly challenging to interpret the clinical importance of rare missense variants in *BRCA1*. Of the missense SNVs assayed, 21.1% (441 out of 2,086) were non-functional (Fig. 2e). Although most remaining missense SNVs were functional (70.6%), there was an enrichment for missense SNVs with intermediate effects (8.1% compared with 4.4% of all other SNVs; Fisher's exact test, $P = 2.7 \times 10^{-6}$).

An advantage of genome editing is that the effect of variants on native regulatory mechanisms such as splicing can be ascertained[22]. Whereas SNVs disrupting canonical splice sites (the two intronic positions immediately flanking each exon) were mostly non-functional (89.5%) or intermediate (5.5%) (Fig. 2e), SNVs positioned 1–3 bp into the exon or 3–8 bp into the intron had variable effects. We defined SNVs in these regions that did not alter the amino acid sequence as 'splice region' variants, of which 22.9% were non-functional (Fig. 2e). SNVs positioned more deeply in introns or in the 5′ untranslated region (UTR) were similar to non-splice-region synonymous SNVs, in that they were much less likely to score as non-functional (intronic, 1.8%; 5′ UTR, 0.0%; and synonymous, 1.3%, as non-functional).

Function scores accurately predict pathogenicity

We next investigated how well our function scores agreed with clinical variant interpretations present in ClinVar. Of 169 SNVs deemed 'pathogenic' in ClinVar that overlapped with our classifications, 162 were designated 'non-functional', two 'functional', and the remaining five 'intermediate'. By contrast, of 22 SNVs deemed 'benign' in ClinVar, 20 were designated 'functional', one 'non-functional', and one 'intermediate' (Fig. 3a). Three SNVs that scored unambiguously discordant with ClinVar suggest potential errors in the available clinical variant interpretations (Supplementary Note 1). A receiver operating

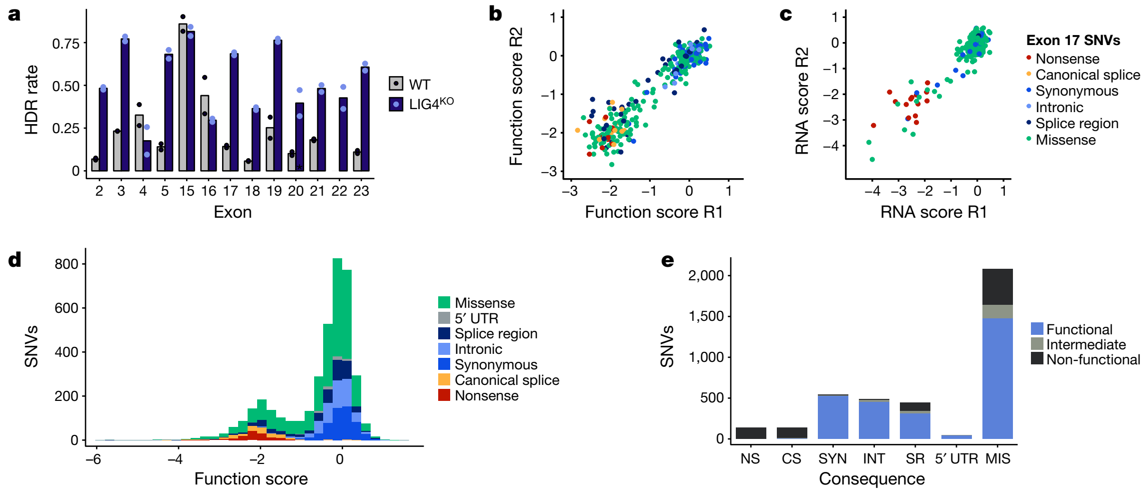


Fig. 2 | Saturation genome editing enables functional classification of 3,893 *BRCA1* SNVs. a, HDR editing rates were calculated for each exon as the fraction of day 5 reads containing the SNV library's fixed synonymous variant (an 'HDR marker' edit). The average of two wild-type HAP1 replicates and two HAP1-LIG4^{KO} replicates is plotted, with points indicating rates for each replicate. (Asterisk denotes missing exon 22 data.) **b**, **c**, Measurements for exon 17 SNVs assayed in HAP1-LIG4^{KO} cells are plotted to show correlations of function scores (**b**, $n = 291$, Spearman's $\rho = 0.88$) and RNA expression scores (**c**, $n = 231$, Spearman's $\rho = 0.61$). Reproducibility is detailed further in Extended Data Fig. 4. **d**, A histogram of 3,893 SNV function scores (averaged from $n = 2$ replicates and normalized across exons) shows how each category of mutation compares to the overall distribution. **e**, The number of SNVs within each category is plotted and coloured by functional classification. (NS, nonsense; CS, canonical splice; SYN, synonymous; INT, intronic; SR, splice region; MIS, missense.)

characteristic (ROC) curve showed a sensitivity of 96.7% at 98.2% specificity when we treat 'likely pathogenic' and 'likely benign' ClinVar annotations as pathogenic and benign, respectively (Fig. 3b). Importantly, sensitivity and specificity are high for missense and splice region SNVs (Extended Data Fig. 7f).

We scored 25.0% (64 out of 256) of VUS and 49.2% (60 out of 122) of SNVs with conflicting interpretations as non-functional (Fig. 3c). Missense VUS from ClinVar were more likely to score as non-functional than missense SNVs that were absent from ClinVar (25.9% compared with 17.2%, Fisher's exact test, $P = 0.002$). Of 3,140 assayed SNVs that were absent from ClinVar, 498 (15.9%) scored as non-functional. The distribution of function scores for the 29 firmly 'pathogenic' missense SNVs confirmed here to be non-functional does not significantly differ from that of the 296 non-functional missense SNVs absent from ClinVar (median −2.05 versus −1.97; Wilcoxon rank-sum test, $P = 0.35$).

We investigated the relationship between our function scores and allele frequencies in large-scale variant databases, such as gnomAD (The Genome Aggregation Database; whole-exome and whole-genome sequencing data from over 120,000 individuals)[27]. Among 302 assayed SNVs that overlap with gnomAD, higher allele frequencies were associated with higher function scores (Extended Data Fig. 8a). For instance, 33 out of 166 (19.9%) of singleton variants were non-functional, whereas only 8 out of 136 (5.9%) non-singleton variants were non-functional (Fisher's exact test, $P = 3 \times 10^{-4}$). A similar trend was observed with the Bravo database (Extended Data Fig. 8b). The FLOSSIES database contains variants observed in around 10,000 women over seventy years old who have not developed breast or ovarian cancer (https://whi.color.com/gene/ENSG00000012048). Of 39 intersecting *BRCA1* SNVs in FLOSSIES, only one scored as non-functional (Extended Data Fig. 8c). Collectively, these observations confirm that *BRCA1* SNVs with higher allele frequencies are more likely to be functional.

Several computational metrics are currently used to the assess deleteriousness of variants and are often included in genetic testing reports. Although our function scores correlate with metrics such as CADD[28], phyloP[29] and Align-GVGD[30], the modesty of these correlations underscores the value of functional assays (Fig. 3d, Extended Data Fig. 9a–g). ROC curve analysis restricted to the 46 missense SNVs deemed 'pathogenic' or 'benign' in ClinVar reveals that SGE function scores outperform these metrics (Extended Data Fig. 9h–l).

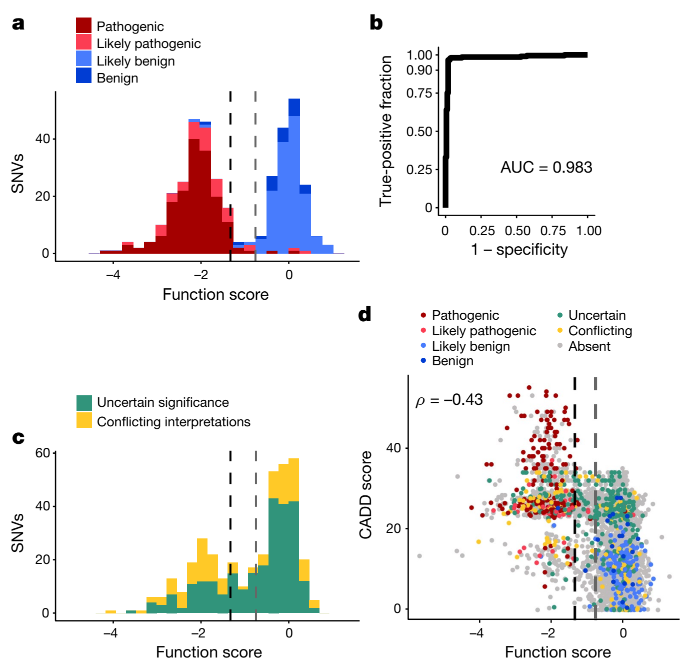


Fig. 3 | SGE function scores are highly accurate at predicting clinical interpretations of *BRCA1* SNVs. a, The distribution of SNV function scores coloured by ClinVar interpretation. Scores are shown for $n = 375$ SNVs with at least a 'one-star' review status in ClinVar and either a 'pathogenic' or 'benign' interpretation (including 'likely'). The dashed lines indicate the functional classification thresholds determined by mixture modelling. Grey divides 'functional' and 'intermediate' (function score = −0.748), and black divides 'intermediate' and 'non-functional' (function score = −1.328). **b**, An ROC curve reveals optimal sensitivity and specificity for classifying the same 375 SNVs in **a** at SGE function score cutoffs from −1.03 to −1.22. **c**, The distribution of scores plotted as in **a** for the 378 SNVs annotated as variants of uncertain significance or with conflicting interpretations. 91.3% of such variants are classified as 'functional' or 'non-functional' by SGE. **d**, CADD scores, which predict deleteriousness, inversely correlate with function scores (Spearman's $\rho = -0.43$, $n = 3,893$ SNVs). SNVs are coloured by ClinVar annotation.

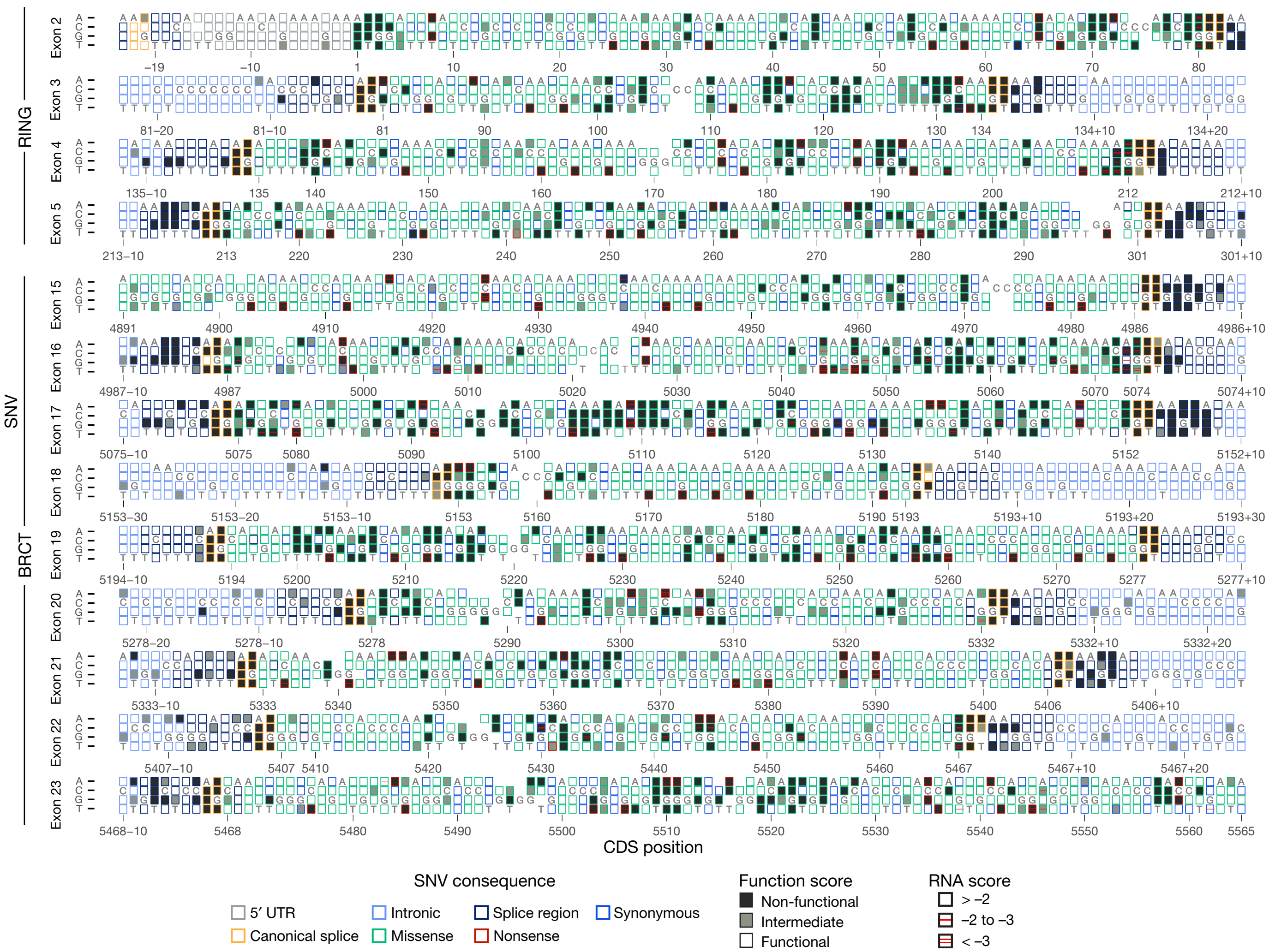


Fig. 4 | Sequence-function maps for 13 *BRCA1* exons. The 3,893 SNVs scored with SGE are each represented by a box corresponding to coding sequence position (NCBI, NM_007294.3) and nucleotide identity. Boxes are filled corresponding to functional classification, and outlined corresponding to the mutational consequence of the SNV. Red lines within boxes mark SNVs depleted in RNA; one line indicates an RNA score between −2 and −3 ($\log_2$ scale) and two lines indicate a score below −3. RNA measurements were determined only for exonic SNVs, excluding exon 18. Reference nucleotides are indicated; blank boxes indicate missing data.

SGE function scores also strongly correlate with the results of assays designed to test particular aspects of BRCA1 activity. For example, they are highly concordant with assays specific for the role of BRCA1 in HDR[9,13,31] and transcriptional activation[12] (Extended Data Fig. 9m, n), as well as with the results of a multiplexed assay that assesses the function of BRCA1 variants in HDR[32].

Mechanisms of *BRCA1* loss-of-function

To gain insights into the various mechanisms by which SNVs in *BRCA1* compromise function, we performed targeted RNA sequencing of *BRCA1* transcripts from edited day 5 cells. We normalized SNV frequencies in cDNA to their frequencies in gDNA to produce mRNA expression scores ('RNA scores') for 96% of the functionally characterized exonic SNVs. Together with function scores, RNA scores enable fine mapping of molecular consequences of SNVs (Fig. 4).

Overall, 89% of non-functional missense SNVs did not reduce RNA levels substantially, suggesting that their effects are mediated at the protein level (Fig. 5a, Supplementary Note 2). Many residues that are sensitive to missense SNVs that do not affect RNA levels map to buried hydrophobic residues or to the zinc-coordinating loops required for RING domain folding[33] (Fig. 5b, c). For example, 20 out of 21 missense SNVs in c.5104–c.5112 were scored as non-functional, including four VUS (Fig. 4). This intolerance to variation is probably due to the hydrophobicity and internal position of Y1703 and F1704, and the polar contacts made between K1702 and a phosphorylated binding partner[34]. This contrasts with a 51-bp stretch spanning exons 21 and 22 (c.5368–c.5418, p.1790–1806) in which none of the 104 missense SNVs assayed were non-functional.

SGE also implicates numerous SNVs that affect expression. For example, all SNVs that disrupt the translation initiation codon score as non-functional, and certain SNVs at the −3, +4 and +5 positions predicted to decrease translational efficiency[35] score as intermediate or non-functional. In addition, 11% of non-functional missense SNVs are depleted from RNA by at least 75%, many of which map to unstructured regions (Fig. 5b, c), suggesting loss-of-function is consequent to reduced mRNA levels rather than disrupted protein function. Consistent with this, the 12 synonymous SNVs classified as non-functional also tended to markedly reduce mRNA levels (median 5.4-fold reduction).

Variants depleted in mRNA probably affect RNA splicing. This is evidenced by an overrepresentation of non-functional exonic SNVs near splice junctions, including low scores for many SNVs at terminal G nucleotides of exons (Fig. 4), non-functional exonic SNVs with low mRNA levels that create new acceptor or donor sequences (Fig. 5d), and the presence of 6–8 bp regions wherein many SNVs have strong effects on mRNA levels, suggestive of exonic splice enhancers[36] (Extended Data Fig. 10a). Certain exons were particularly prone to harbour non-functional SNVs with low RNA scores. In exon 16, for

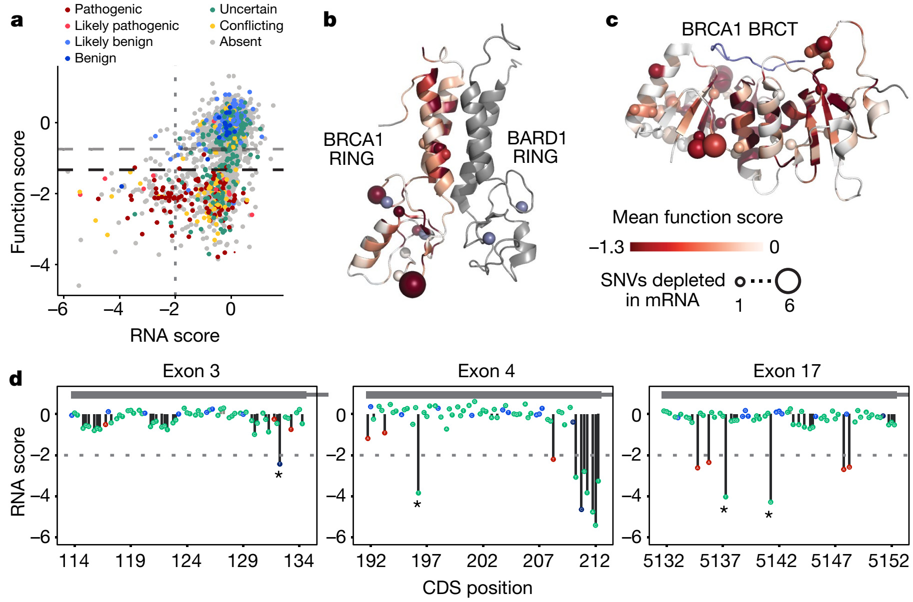


Fig. 5 | Measuring SNV mRNA abundance and function in parallel delineates mechanisms of variant effect. a, Function scores are plotted against RNA scores for all exonic synonymous and missense SNVs scored ($n = 2{,}646$). Horizontal dashed lines indicate functional thresholds, and the vertical dotted line marks an RNA score of −2. **b**, **c**, Function scores for all SNVs were mapped onto the structures of the RING (**b**, PDB 1JM7) and BRCT (**c**, PDB 1T29) domains in shades of red by averaging missense SNV scores at each amino acid position. The number of SNVs that cause more than 75% reduction in mRNA levels at each amino acid position is represented by the size of the sphere at the alpha-carbon of each residue. Grey denotes residues not assayed and the BACH1 peptide bound to the BRCT structure is coloured blue. **d**, SNV RNA scores are plotted by transcript position, with lines to the x axis denoting SNV functional classifications (no line, functional; grey line, intermediate; black line, non-functional; SNVs coloured by consequence as in Fig. 2c). The horizontal dashed line in each plot marks an RNA score of −2, corresponding to 75% reduction in mRNA. Examples of non-functional SNVs with low RNA scores that create new 5′-GU splice donor motifs are indicated with asterisks.

instance, 46 of 244 SNVs (excluding nonsense) were non-functional (Extended Data Fig. 10a). Most of these (26 out of 46) reduced RNA levels by ≥2-fold, and 15 by ≥4-fold. By contrast, in exon 19, 55 of 234 SNVs (excluding nonsense) were non-functional, but none lowered expression by ≥2-fold (Extended Data Fig. 10b). Exon 19 also completely lacks non-functional SNVs in its flanking intronic regions (apart from the acceptor and donor sites), suggesting it is robustly spliced.

Discussion

Here we applied SGE to critical domains of *BRCA1*, characterizing the consequences of nearly 4,000 SNVs in their native genomic context and obtaining a bimodal distribution of functional effects. A benefit of functional data is that measurements are systematically derived, independent of prior expectation[37]. Because we measured cell survival, the effects of SNVs on multiple layers of gene function (for example, splicing, translation, and protein activity) are effectively integrated. Our study has several caveats (Supplementary Note 3), most notably that we used a survival assay in HAP1 cells as opposed to a more physiologically appropriate model. However, our data are validated by high concordance with the available evidence for clinical pathogenicity.

High sensitivity and specificity were obtained for both missense and splice region SNVs, the classes of variants that are most problematic for clinical interpretation. Our review of firmly discordant SNVs suggests that our true accuracy may be higher than calculated using ClinVar assertions as a gold standard (Supplementary Note 1). These discordances highlight the importance of integrating new evidence as it becomes available and updating databases accordingly. For instance, the submissions in the Breast Cancer Information Core, which mostly date to the early 2000s, underlie 51 conflicting interpretations. SGE scores support the more recent classification in the vast majority of such conflicts (Supplementary Table 2).

The interpretation of genetic variation is presently the rate-limiting step for genomic medicine. The fact that more than 70% of ClinVar variants and more than 95% of non-ClinVar variants assayed here have never been observed in more than 120,000 humans represented in gnomAD illustrates the challenges facing observational approaches to variant interpretation. Given this, a pressing question is how best to integrate functional data into existing clinical variant classification schemes[38]. The predictive power demonstrated here suggests that SGE function scores classify variants with more than 95% accuracy. As current standards for defining 'likely' pathogenic and benign variants accept comparable uncertainty[39], we argue that a failure to incorporate function scores would be a missed opportunity.

Optimal weighting of different approaches might further improve classification of variants lacking genetic evidence. For unexpected functional classifications, such as synonymous SNVs with low scores, and for cases in which the clinical evidence is contradictory, functional data can provide specific hypotheses to test. For example, c.5044G>A, for which our data contradicts ClinVar, could be disambiguated by testing *BRCA1* mRNA levels in individuals carrying this SNV. The approximately 6% of SNVs exhibiting intermediate function scores remain beyond definitive interpretation. The fact that we observe an excess of missense SNVs with intermediate scores suggests that some of these may be hypomorphic *BRCA1* alleles[40]. Further studies will be necessary to assess the risk conferred by these variants.

We prioritized the RING and BRCT domains, but SGE of all exons of *BRCA1* is justified, and the essentiality of *BRCA2*, *PALB2*, *BARD1* and *RAD51C* in HAP1 cells suggests that these genes are assayable by the same method. For other genes, assays compatible with saturation genome editing (for example, drug selection, FACS on phenotypic markers) may need to be developed and validated. Scaling SGE to many loci also promises to improve our understanding of how diverse biological functions are encoded by the genome.

Here we show that SGE is a viable strategy for functionally classifying thousands of variants in a clinically actionable gene, most of which have yet to be observed in a human. We anticipate function scores will prove valuable, both for adjudicating hundreds of observed *BRCA1* variants for which the interpretation is currently ambiguous, as well as

for providing immediate functional assessments for newly observed variants. This work may also serve as a blueprint for the comprehensive functional analysis of all potential SNVs in clinically actionable genes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0461-z.

Received: 29 January 2018; Accepted: 26 July 2018;
Published online 12 September 2018.

1. Rehm, H. L. et al. ClinGen–the Clinical Genome Resource. *N. Engl. J. Med*. **372**, 2235–2242 (2015).
2. Kuchenbaecker, K. B. et al. Risks of breast, ovarian, and contralateral breast cancer for *BRCA1* and *BRCA2* mutation carriers. *J. Am. Med. Assoc*. **317**, 2402–2416 (2017).
3. Hall, J. M. et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
4. Olopade, O. I. & Artioli, G. Efficacy of risk-reducing salpingo-oophorectomy in women with *BRCA-1* and *BRCA-2* mutations. *Breast J*. **10**, S5–S9 (2004).
5. Rebbeck, T. R. et al. Bilateral prophylactic mastectomy reduces breast cancer risk in *BRCA1* and *BRCA2* mutation carriers: the PROSE Study Group. *J. Clin. Oncol*. **22**, 1055–1062 (2004).
6. Easton, D. F. et al. Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med*. **372**, 2243–2257 (2015).
7. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. **44**, D862–D868 (2016).
8. Millot, G. A. et al. A guide for functional analysis of BRCA1 variants of uncertain significance. *Hum. Mutat*. **33**, 1526–1537 (2012).
9. Ransburgh, D. J. R., Chiba, N., Ishioka, C., Toland, A. E. & Parvin, J. D. Identification of breast tumor mutations in *BRCA1* that abolish its function in homologous DNA recombination. *Cancer Res*. **70**, 988–995 (2010).
10. Pierce, A. J., Hu, P., Han, M., Ellis, N. & Jasin, M. Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mammalian cells. *Genes Dev*. **15**, 3237–3242 (2001).
11. Bouwman, P. et al. A high-throughput functional complementation assay for classification of *BRCA1* missense variants. *Cancer Discov*. **3**, 1142–1155 (2013).
12. Woods, N. T. et al. Functional assays provide a robust tool for the clinical annotation of genetic variants of uncertain significance. *NPJ Genom. Med.* **1**, 16001 (2016).
13. Starita, L. M. et al. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).
14. Steffensen, A. Y. et al. Functional characterization of *BRCA1* gene variants by mini-gene splicing assay. *Eur. J. Hum. Genet*. **22**, 1362–1368 (2014).
15. de la Hoya, M. et al. Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. *Hum. Mol. Genet*. **25**, 2256–2268 (2016).
16. Ghosh, R., Oak, N. & Plon, S. E. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol*. **18**, 225 (2017).
17. Gibson, T. J., Seiler, M. & Veitia, R. A. The transience of transient overexpression. *Nat. Methods* **10**, 715–721 (2013).
18. Moynahan, M. E., Chiu, J. W., Koller, B. H. & Jasin, M. BRCA1 controls homology-directed DNA repair. *Mol. Cell* **4**, 511–518 (1999).
19. Drost, R. et al. BRCA1 RING function is essential for tumor suppression but dispensable for therapy resistance. *Cancer Cell* **20**, 797–809 (2011).
20. Shakya, R. et al. BRCA1 tumor suppression depends on BRCT phosphoprotein binding, but not its E3 ligase activity. *Science* **334**, 525–528 (2011).
21. Vega, A. et al. The R71G BRCA1 is a founder Spanish mutation and leads to aberrant splicing of the transcript. *Hum. Mutat*. **17**, 520–521 (2001).
22. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).
23. Blomen, V. A. et al. Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).
24. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protoc*. **8**, 2281–2308 (2013).
25. Beumer, K. J. et al. Efficient gene targeting in *Drosophila* by direct embryo injection with zinc-finger nucleases. *Proc. Natl Acad. Sci. USA* **105**, 19821–19826 (2008).
26. Essletzbichler, P. et al. Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. *Genome Res*. **24**, 2059–2065 (2014).
27. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
28. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet*. **46**, 310–315 (2014).
29. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. **20**, 110–121 (2010).
30. Tavtigian, S. V., Byrnes, G. B., Goldgar, D. E. & Thomas, A. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Hum. Mutat*. **29**, 1342–1354 (2008).
31. Towler, W. I. et al. Analysis of BRCA1 variants in double-strand break repair by homologous recombination and single-strand annealing. *Hum. Mutat*. **34**, 439–445 (2013).
32. Starita, L. M. et al. A multiplexed homology-directed DNA repair assay reveals the impact of over 1,000 BRCA1 missense substitution variants on protein function. *Am. J. Hum. Genet*. https://doi.org/10.1016/j.ajhg.2018.07.016 (2018).
33. Brzovic, P. S., Rajagopal, P., Hoyt, D. W., King, M. C. & Klevit, R. E. Structure of a BRCA1–BARD1 heterodimeric RING–RING complex. *Nat. Struct. Biol*. **8**, 833–837 (2001).
34. Shiozaki, E. N., Gu, L., Yan, N. & Shi, Y. Structure of the BRCT repeats of BRCA1 bound to a BACH1 phosphopeptide: implications for signaling. *Mol. Cell* **14**, 405–412 (2004).
35. Wegrzyn, J. L., Drudge, T. M., Valafar, F. & Hook, V. Bioinformatic analyses of mammalian 5′-UTR sequence properties of mRNAs predicts alternative translation initiation sites. *BMC Bioinformatics* **9**, 232 (2008).
36. Desmet, F.-O. et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. **37**, e67 (2009).
37. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
38. Starita, L. M. et al. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet*. **101**, 315–325 (2017).
39. Plon, S. E. et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat*. **29**, 1282–1291 (2008).
40. Lovelock, P. K. et al. Identification of BRCA1 missense substitutions that confer partial functional activity: potential moderate risk variants? *Breast Cancer Res*. **9**, R82 (2007).

Acknowledgements We thank M. Spielmann, D. Witten, A. McKenna, M. Kircher, M. Dougherty, J. Lazar, Y. Yin, and B. Shirts for insights on data analysis and/or comments on the manuscript, J. Kitzman for sharing reagents and protocols, R. Acuña-Hidalgo, J. Milbank, and E. van Veen for experimental assistance, and the Feng Zhang laboratory for sharing Cas9/gRNA plasmids. This work was supported by the Brotman Baty Institute for Precision Medicine, an NIH Director's Pioneer Award (DP1HG007811 to J.S.) and a training award from the National Cancer Institute (F30CA213728 to GMF). J.S. is an Investigator of the Howard Hughes Medical Institute.

Reviewer information *Nature* thanks H. Rehm, J. Weissman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.M.F., J.S. and L.M.S. conceived the project. G.M.F. designed experiments. G.M.F. and R.M.D. performed experiments with assistance from B.M., M.D.Z., A.P.L., L.M.S. and M.G. G.M.F. performed analysis with assistance from L.M.S., J.D.J., X.H. and R.M.D. G.M.F, J.S. and L.M.S. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0461-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0461-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.M.S. or J.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized; the investigators were not blinded to allocation during experiments and outcome assessment.

HDR pathway essentiality analysis in HAP1 cells. HAP1 cells were derived from KBM7 cells (a near-haploid immortalized chronic myelogenous leukaemia line) by introduction of induced pluripotent stem cell factors[41]. *HAP1* gene essentiality scores were obtained[23] and filtered on genes with more than 20 mapped gene-trap insertions ($n = 14{,}306$). Of 78 HDR genes defined by the Gene Ontology term 'double-strand break repair via homologous recombination' (GO:0000724), 66 were among the 14,306 genes included in analysis. To rank genes by essentiality, they were first ordered by q value (low to high) and second by the proportion of gene-trap insertions in the sense orientation (low to high). HDR pathway genes implicated in cancer (labelled in Fig. 1a) were defined as those included on the University of Washington BROCA sequencing panel[42].

gRNA design and cloning. All CRISPR gRNAs used in SGE and essentiality experiments were cloned into pX459[24]. This plasmid expresses the gRNA from a U6 promoter, as well as a Cas9-2A-puromycin resistance (-puroR) cassette. *S. pyogenes* Cas9 target sites were chosen for SGE experiments on multiple criteria, assessed in the following order: (i) to induce cleavage within *BRCA1* coding sequence, (ii) to target a genomic site permissive to synonymous substitution within the guanine dinucleotide of the PAM or the protospacer, (iii) to have minimal predicted off-target activity[43], (iv) to have maximal predicted on-target activity[44].

Complementary oligonucleotides ordered from Integrated DNA Technologies (IDT) were annealed, phosphorylated, diluted and ligated into BbsI-digested and gel-purified pX459, as described previously[24]. Ligation reactions were transformed into *Escherichia coli* (Stellar competent cells, Takara), which were plated on ampicillin. Colonies were cultured and Sanger-sequenced to confirm correct gRNA sequences. Purification of sequence-verified plasmids for transfection was performed with the ZymoPure Maxiprep kit (ZymoResearch). For targeting *LIG4* in HAP1 cells, pX458[24] was used instead of pX459, which expresses EGFP in lieu of puroR.

HDR library design and cloning. Array-synthesized oligonucleotides were designed as follows for each saturation genome editing region (that is, a *BRCA1* exon). The sequence to be mutated (~100 bp) was obtained from the human genome (hg19) and a synonymous substitution was introduced at the chosen Cas9 target site (for example, a substitution at the PAM site). This 'fixed' substitution in the library was included in design to serve multiple purposes: (i) plasmid library molecules harbouring the substitution are predicted to be cleaved less frequently by Cas9–gRNA complexes, (ii) SNVs introduced to cells are predicted to be depleted via Cas9 re-cutting less frequently as a consequence of the fixed substitution, and (iii) sequencing reads can be filtered on the fixed substitution to distinguish true SNVs introduced via HDR from sequencing errors. A second synonymous substitution at an alternative CRISPR target site was introduced to the sequence as well, such that the SNV library for each exon would be compatible with multiple gRNAs. Next, a sequence was created for every single nucleotide substitution on this template. For all sequences, adapters were added to both ends to enable PCR amplification from the oligonucleotide pool. For each SGE region, the total number of oligonucleotides designed was three times the length of the region, plus the oligonucleotide template without any SNV (for example, for a 100-bp SGE region, 301 total oligonucleotides were designed).

Pooled oligonucleotides were synthesized (Agilent Technologies). Primers designed to amplify the subset of oligonucleotides corresponding to a single region of an exon were used to perform PCR with Kapa HiFi Hot-start Ready Mix (Kapa HiFi, Kapa Biosystems). PCR products were purified with Ampure beads (Agencourt) to be used in subsequent library cloning reactions.

Homology arms were cloned into pUC19 by PCR-amplifying (Kapa HiFi) regions surrounding each targeted exon from HAP1 gDNA. Primers for these reactions were designed such that homology arms would be between 600 bp and 1,000 bp on both sides of the targeted region. Adapters homologous to pUC19 were added to primers to facilitate NEBuilder HiFi Assembly cloning (NEB) into a linearized pUC19 vector. Cloning reactions were transformed into Stellar competent cells and selected with ampicillin. Plasmid DNA was isolated from colonies (Qiagen MiniPrep kit) and sequence-verified.

To construct the HDR library, homology arm plasmids were linearized via PCR using primers that conferred 15–20 bp of terminal overlap with the adapter sequences flanking each PCR-amplified oligonucleotide pool. This sequence overlap enabled cloning via the NEBuilder HiFi Assembly Cloning Kit (NEB). Cloning reactions were transformed into Stellar competent cells, and a small proportion (1%) of the transformation was plated on ampicillin-containing plates to assess efficiency. All remaining transformed cells were grown directly in 100 ml of medium with ampicillin for 16–18 h, and plasmid DNA from the culture was isolated (ZymoPure Maxiprep kit) to produce each final HDR library.

HAP1 cell culture. Quality-controlled wild-type HAP1 cells were purchased (Haplogen/Horizon Discovery) and cultured in medium comprising Iscove's Modified Dulbecco's Medium (IMDM) with L-glutamine and 25 mM HEPES (GIBCO) supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals) and 1% penicillin–streptomycin (GIBCO). Cells were grown on plates at 37 °C with 5% CO_2, and passaged before becoming confluent. For routine passaging, cells were washed once with 1× phosphate-buffered saline (PBS, Gibco), trypsinized with 0.25% trypsin with EDTA (Gibco), resuspended in medium, centrifuged for 5 min at 300g, and then resuspended and plated.

A monoclonal *LIG4*-knockout HAP1 line (HAP1-LIG4^{KO}) was generated by transfecting a plasmid expressing a Cas9-2A-GFP cassette and a gRNA targeting the human *LIG4* coding sequence (gRNA sequence: 5′-GCATAATGT CACTACAGATC-3′) into wild-type HAP1 cells. Single GFP-expressing HAP1 cells were sorted into wells of a 96-well plate and cultured. After two weeks, gDNA was collected and Sanger sequencing was performed to assess *LIG4* editing. A clone with a 4-bp deletion was identified and expanded further for use in saturation genome editing experiments.

HAP1 cells can spontaneously revert to a diploid state in cell culture. Therefore, to sort a 1n-enriched population of cells before transfection, cells were stained for DNA content with Hoechst 34580 (BD Biosciences) at 5 μg ml^{-1} medium for 1 h at 37 °C. FACS was performed to isolate 1–2 × 10^6 cells from the lowest intensity Hoechst peak, corresponding to 1n ploidy. These cells were expanded for seven days before transfection.

Transfection of HAP1 cells. For all experiments, HAP1 cells were transfected using TurboFectin 8.0 (Origene) according to manufacturer's protocol. A 2.5× volume of Turbofectin was added to the transfection mix for each μg of plasmid DNA in Opti-Mem (Life Technologies). For each SGE transfection, 10 million cells were passaged to a 10-cm dish. The next day (day 0), cells were co-transfected with 12 μg of the Cas9/gRNA plasmid (pX459) and 3 μg of the SNV library corresponding to a single exon. Negative control transfections were performed for each library using a pX459 vector targeting *HPRT1* instead of *BRCA1*, thus preventing genomic integration of the library. On day 1, cells were passaged into medium supplemented with puromycin (1 μg ml^{-1}) to select for successfully transfected cells. On day 4, cells were washed twice and passaged to 6-cm plates in regular media.

Cell populations were sampled on day 5 and day 11 for all SGE experiments. On day 5, half of the cells were pelleted and frozen and the other half passaged. The cells were passaged on day 8 into 15-cm dishes and then harvested on day 11. Negative control transfections were harvested on day 5 and used to confirm that PCR amplicons were not derived from the plasmid DNA of the SNV library.

For the luminescence-based viability assay, HAP1 cells were plated at 35–40% confluency in a 6-well dish (approximately 1.2 million cells per well per target) then transfected with 1.5 μg Cas9/gRNA plasmid targeting coding exons of HDR genes or controls the following day. After 24 h of transfection, the cells were plated in time-point triplicates at 20,000 cells per well in 96-well clear bottom plates in medium with and without puromycin. Cells without puromycin were assessed 4 h after plating to establish baseline absorbance for each target. Cell survival was assessed at day 2, day 5, and day 7 after transfection using the CellTiterGlow reagent (Promega, 1:10 dilution of suggested reagent). Luminescence at 135-nm absorbance was measured using a Synergy plate reader (Biotek Instruments).

Nucleic acid sampling and sequencing library production. For obtaining wild-type HAP1 genomic DNA for cloning homology arms and for genotyping the HAP1-LIG4^{KO} cell line, DNA was isolated using the DNeasy kit (Qiagen). For each SGE experiment, DNA and total RNA were purified using the AllPrep kit (Qiagen). DNA samples were quantified with the Qubit dsDNA Broad Range kit (Thermo Fisher) and RNA samples by UV spectrometry (Nanodrop). PCR primers for genomic DNA were designed such that one primer would anneal outside of the homology arm sequence, thereby selecting for amplicons derived from gDNA and not plasmid DNA. PCR conditions were optimized using gradient qPCR on wild-type *HAP1* gDNA.

All gDNA collected from the population of day-5 cells was sampled by performing many PCR reactions in parallel on a 96-well plate, using 250 ng of gDNA per 50 μl reaction such that all day-5 gDNA was used in PCR (Kapa HiFi). At least as many PCR reactions were performed for day-11 samples (which yielded more gDNA) to ensure adequate sampling. PCRs were performed for the minimal number of cycles needed to complete amplification, with cycling conditions as specified in the Kapa HiFi protocol. An additional PCR was performed using day-5 gDNA from negative control transfections for each exon.

After PCR, multiple wells of amplicons from the same sample were pooled and purified using Ampure beads. Next, a nested qPCR was performed using the first reaction as template to produce a smaller amplicon with custom sequencing adapters ('PU1L' and 'PU1R'), which was likewise purified with Ampure beads. The SGE libraries were also PCR-amplified at this step, starting from 50 ng of plasmid DNA. Lastly, a final qPCR was performed using purified products from the second reaction as template to add dual sample indexes and flow cell adapters.

RNA was sampled from day-5 HAP1-LIG4^{KO} cells (AllPrep, Qiagen). Reverse transcription followed by RNase H treatment was performed on all collected RNA or a maximum of 5 μg per sample (Superscript IV Kit, Life Technologies). This reaction was primed with a gene-specific primer complementary to the 3′ UTR in exon 23 of *BRCA1*. Primers were designed for each exon to amplify across exon junctions, and reaction conditions were optimized using gradient PCR. cDNA was distributed into five equal PCR reactions, which were run on a qPCR machine and then pooled in equal ratios. Flow cell adapters and sample indexes were added in an additional reaction (as for gDNA samples).

All sequencing libraries were purified with Ampure beads, quantified with the Qubit dsDNA High Sensitivity kit (Life Technologies), diluted and denatured for sequencing in accordance with protocols for the Illumina NextSeq or MiSeq machines.

Sequencing and data analysis. Sequencing was performed on an Illumina NextSeq or MiSeq instrument, allocating about 3 million reads to each gDNA and cDNA sample, 1 million reads for each HDR library, and 500,000 reads for each negative control sample. gDNA samples for individual exons were sequenced on the same run. In total, 300 cycle kits were used, with 150 cycles for read 1 and read 2 each, and 19 cycles for dual index reads. Custom sequencing primers and indexing primers are provided in Supplementary Table 3. Illumina PhiX control DNA was added to each sequencing run (around 10% MiSeq, around 30–40% NextSeq) to improve base calling.

We used bcl2fastq 2.16 (Illumina) to call bases and perform sample demultiplexing and fastqc 0.11.3 was run on all samples to assess sequencing quality. SeqPrep was used with the following parameters to perform adapter trimming and to merge perfectly matched overlapping read pairs: '-A GGTTTGGAGCGAGATTGATAAAGT -B CTGAGCTCTCTCACAGCCATTTAG -M 0.1 -m 0.001 -q 20 -o 20'. Merged reads containing 'N' bases were removed. Reads from cDNA samples were removed if they contained indels or did not perfectly match transcript sequence flanking each targeted exon. Remaining cDNA reads were processed to match genomic DNA amplicons by removing flanking exonic sequence and replacing it with the exon's corresponding intronic sequence. All reads were then aligned to reference gDNA amplicons for each exon using the needleall command in the EMBOSS 6.4.0 package with the following parameters: '-gapopen 10 -gapextend 0.5 -aformat sam'. Reads not aligning to the reference amplicon (alignment score, <300) were removed from analysis. To analyse indels, unique cigar counts were quantified from day-5 and day-11 samples using a custom Python script. Reads were classified as HDR events for rate calculations if the programmed edit or edits to the PAM or protospacer (HDR marker edits) were observed in the alignment. Variants without identifiable markers of HDR were not used. Abundances of SNVs were quantified only from aligned reads that had no other mismatches or indels, with the exception of the HDR markers. SNV reads with only the cut-site proximal HDR marker were summed with reads that had both HDR markers to get total abundances for each SNV in each sample, to which a pseudocount of 1 was added to all variants present in either the library, day-5 or day-11 sample. Frequencies for each SNV were calculated as SNV reads over total reads. SNV measurements from wild-type HAP1 cells and HAP1-LIG4^{KO} cells were processed separately at all steps.

Specific exon 2 splice junctions were queried by counting the number of reads from cDNA samples that perfectly matched specific isoform junctions. Two 14-bp sequences spanning the end of exon 1 and the beginning of exon 2 were counted to measure use of the canonical junction (5′-TCTGGTTCATTGGA-3′ and 5′-TCTGGTTCACTGGA-3′; the latter of which contains an HDR marker introduced during editing). The 14-bp sequence spanning the end of exon 1 and the portion of exon 2 corresponding to the reported alternative AG acceptor site[45,46] was (5′-TAAAGAAAGAAATG-3′). The proportion of the total reads counted containing the latter sequence was used to approximate the relative contribution of the alternative acceptor site.

Modelling positional biases of library integration. Positional biases in editing rates were modelled for each SNV by using a LOESS regression to fit the $\log_2$ day 5 over library ratios as a function of chromosomal position. To avoid modelling biological effects instead of positional effects, the model was fit only on the subset of SNVs that were not substantially depleted between any two time points in the experiment (that is, SNVs with day 5 over library ratios greater than 0.5 and day 11 over day 5 ratios greater than 0.8.). The regression was performed for each exon replicate, using the 'loess' function in R with span = 0.15. Each model was extended flatly outward to include any positions not fit (a total of 22 nucleotides of sequence on the edges of the edited regions). We subtracted positional fit (the model's output) for each SNV from its $\log_2$ day 11 over library ratio to get position-adjusted ratios for each SNV.

Normalizing scores within and across exons. Position-adjusted $\log_2$ day 11 over library ratios were normalized first across exon replicates, and then across all assayed exons. To do this, scores from within each replicate were linearly scaled such that the median synonymous and median nonsense SNVs within the replicate would match the median synonymous and median nonsense SNV values averaged across replicate experiments. The ensuing SNV scores for each replicate were then normalized across all exons in the same manner, such that each exon's median synonymous and median nonsense SNV scores would match the global median synonymous and the global median nonsense SNV scores, respectively.

SNV functional class assignment. Function scores were averaged across replicates and a mixture model was used to estimate the probability that each SNV's score was drawn from the non-functional distribution of scores. The non-functional distribution was defined as nonsense SNVs across all exons. The functional distribution was defined as exonic synonymous SNVs not within 3 bp of splice junctions and with RNA scores within 1 standard deviation of the median synonymous SNV. This definition does not fully guarantee that these SNVs have no functional consequence. The means and variances of the 'non-functional' and 'functional' groups were fixed and a model was fit using the normalmixEM function of the mixtools package in R, with starting component proportions set to 0.5. The posterior probabilities generated from the model were used as point estimates of the probability of drawing each SNV's score from the non-functional distribution (P_{nf}). Functional classifications were made by setting thresholds for P_{nf} as follows: $P_{nf} > 0.99$ = 'non-functional', $0.01 < P_{nf} < 0.99$ = 'intermediate', $P_{nf} < 0.01$ = 'functional'.

Independent of mixture modelling, ROC curves were used to assess performance of SGE data and other metrics' ability to predict assigned ClinVar classifications. These analyses were performed with the 'plotROC' package in R, and Youden's *J*-statistic (sensitivity plus specificity minus 1) was calculated to determine optimal values reported in text.

Variant filtering. A small minority of SNVs that could not be accurately scored were removed from analysis. If a SNV was not present in the HDR library at a frequency over 1 in 10^4, it was presumed to have been lost in oligonucleotide synthesis or cloning and was removed. Further, if a SNV was not observed with complete HDR markers at a frequency over over 1 in 10^5 in day-5 genomic DNA samples from both replicate experiments, it was removed. SNVs introduced near the CRISPR recognition site have the potential to facilitate Cas9 re-cutting of the locus (for example, by replacing the PAM edit or introducing an alternative PAM site). Because these SNVs are likely to score lower consequent to Cas9 editing biases and not their effects on gene function, SNVs were filtered that created increased potential for re-cutting as follows: When an HDR marker mutation used to disrupt editing occurred at position 2 of the PAM (for example, 'NGG' to 'NCG'), SNVs that replaced this marker with an alternate base were removed to prevent biases introduced by re-cutting non-canonical *S. pyogenes* Cas9 PAMs (for example, 'NAG', 'NTG'). Additionally, variants that created a new PAM 1 bp 3′ of the mutated PAM were excluded owing to the potential for re-cutting (for example, unedited PAM: 5′-<u>NGG</u>A, edited PAM with HDR marker: 5′-<u>NCG</u>A, filtered out SNV that creates new PAM +1 bp 3′: 5′-N<u>CGG</u>). (Extended Data Figure 6 describes re-cutting observed at alternative PAMs.) To prevent misinterpretation, we also removed SNVs that created amino acid changes specific to the context of the library's fixed edits (for example, if in the unedited background, the SNV causes an *X* to *Y* change, but with a fixed edit in the same codon, the SNV causes an *X* to *Z* change). We also applied this logic to remove SNVs that introduced splice donor sites only in the context of the edited PAM, and SNVs that create splice donor sites in the unedited context but not in the context of the edited PAM.

The RNA scores for exon 18 samples were neither well-correlated across replicates nor with SNV abundances in genomic DNA, indicating probable bottlenecking in library preparation. Therefore, RNA data from exon 18 was excluded. Wild-type HAP1 function scores from exon 22 were excluded because there was an unusually high correlation between SNV frequencies sampled from the plasmid library and from day-5 gDNA, suggesting plasmid contamination in gDNA sequencing. This problem was fixed by designing a new primer to prepare gDNA sequencing samples from HAP1-LIG4^{KO} cells.

External data sources. Variant annotations were downloaded from CADD[28] version 1.3 (http://cadd.gs.washington.edu/download). This included the following scores: mammalian phyloP, Grantham deviation, SIFT, Polyphen-2 and CADD. Align-GVGD scores were obtained by running the Align-GVGD program on *BRCA1* sequences conserved to sea urchin. ClinVar data were downloaded on 2 January 2018 for all germline SNVs with at least a 1-star annotation. SNVs annotated as 'Benign/Likely benign' were grouped with 'Likely benign' SNVs and SNVs classified 'Pathogenic/Likely pathogenic' were grouped with 'Likely pathogenic' SNVs. SNV allele frequencies were obtained from http://gnomad.broadinstitute.org/ on 26 December 2017 for gnomAD[27], from https://bravo.sph.umich.edu/freeze5/hg38/ on 19 November 2017 for Bravo, and from https://whi.color.com/ on 9 October 2017 for FLOSSIES data. The hg19 UCSC Genome Browser was accessed from https://genome.ucsc.edu/ on 1 May 2018 for chr17:41,276,108–41,276,139. Throughout this study, *BRCA1* exons, coding nucleotide positions, and amino acid positions are referenced by the ClinVar transcript annotation for *BRCA1*, NCBI transcript NM_007294.3.

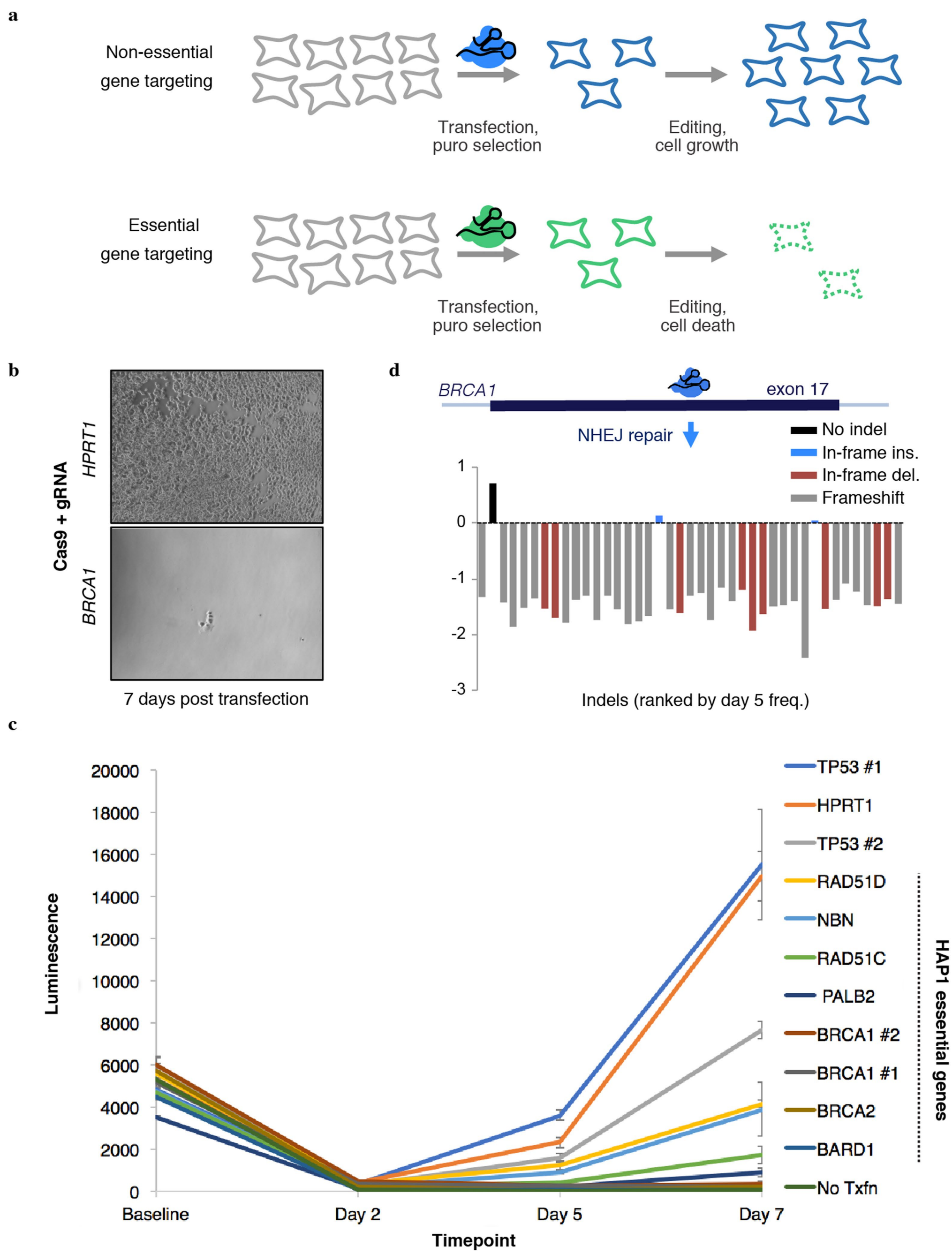
Statistical reporting. All statistical tests described were performed as two-tailed tests using the R software package.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

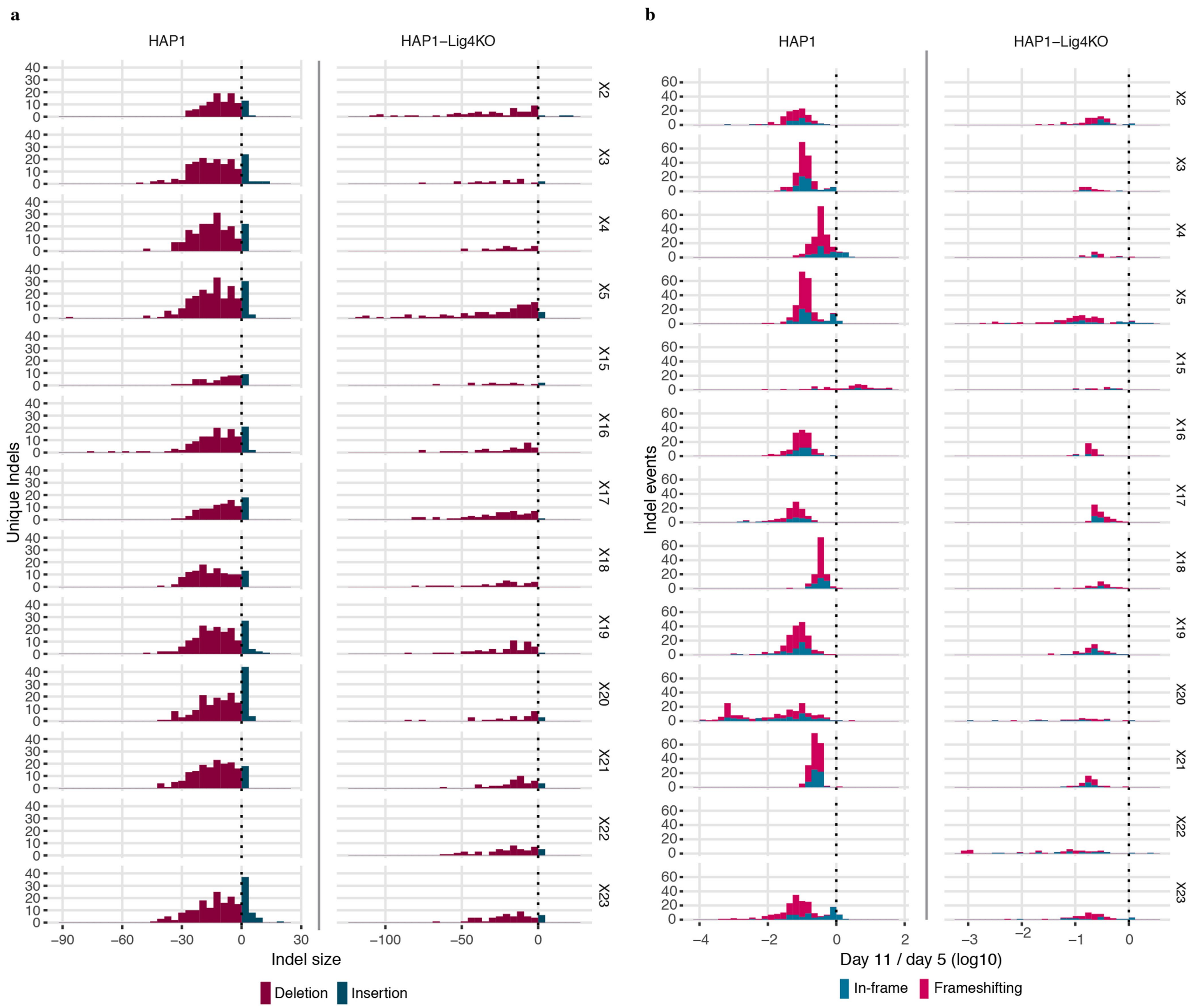
Code availability. Custom scripts for analysing sequencing data were written in Python and R. All code is available at: https://github.com/shendurelab/saturationGenomeEditing_pipeline.

Data availability. Function scores are freely available for all non-profit uses (see https://sge.gs.washington.edu/BRCA1/), as well as by non-exclusive license under reasonable terms to commercial entities that have committed to open sharing of *BRCA1* sequence variants. Sequencing data are available at Gene Expression Omnibus under accession GSE117159.

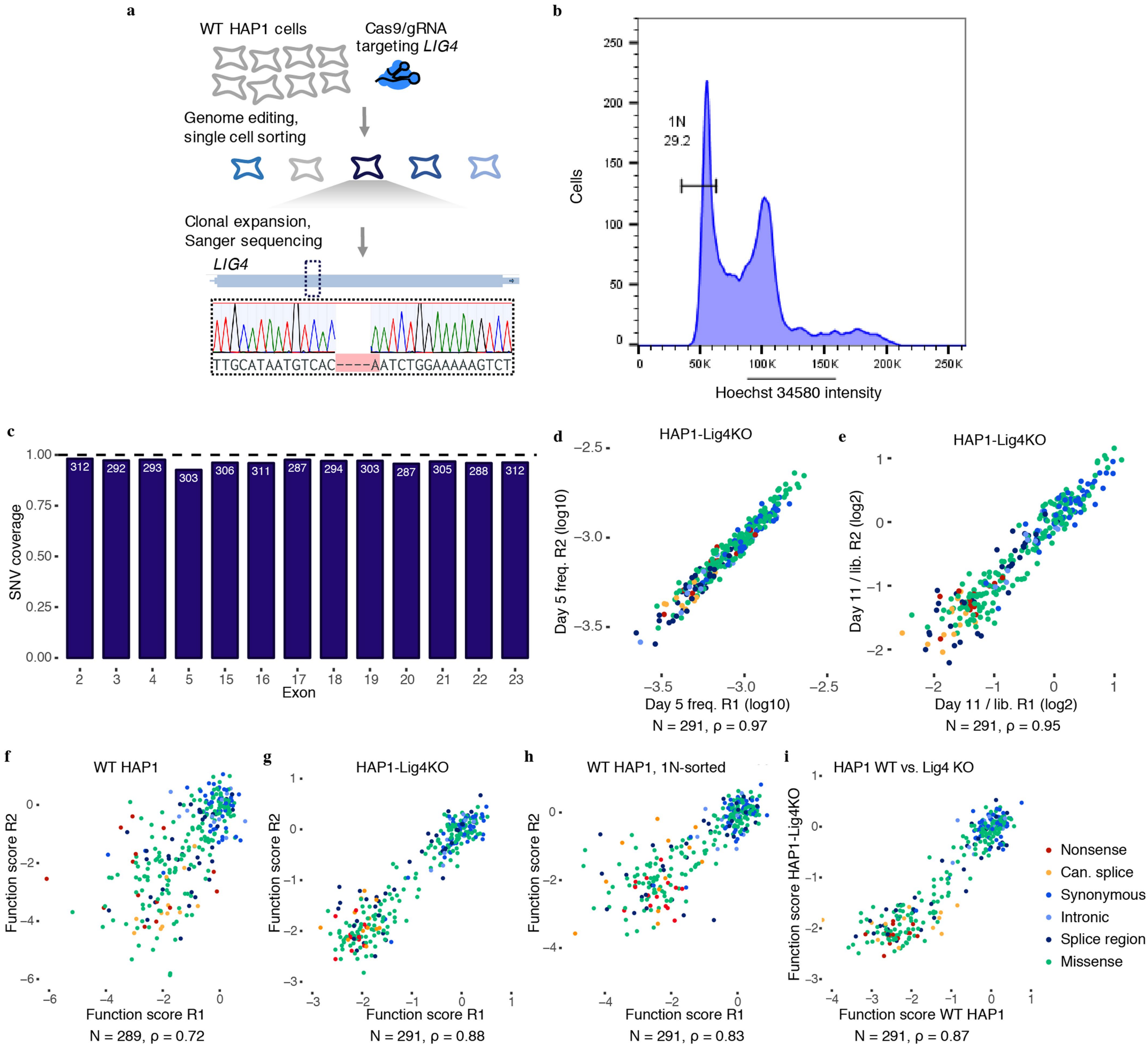
41. Carette, J. E. et al. Ebola virus entry requires the cholesterol transporter Niemann–Pick C1. *Nature* **477**, 340–343 (2011).
42. Walsh, T. et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **107**, 12629–12633 (2010).
43. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
44. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
45. Colombo, M. et al. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the *BRCA1* locus: a report from the ENIGMA consortium. *Hum. Mol. Genet.* **23**, 3666–3680 (2014).
46. Romero, A. et al. BRCA1 alternative splicing landscape in breast tissue samples. *BMC Cancer* **15**, 219 (2015).
47. Tavtigian, S. V. et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* **43**, 295–305 (2006).
48. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
49. Adzhubei, I. & Jordan, D. M. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Gen.* **76**, 7.20.1–7.20.41 (2013).



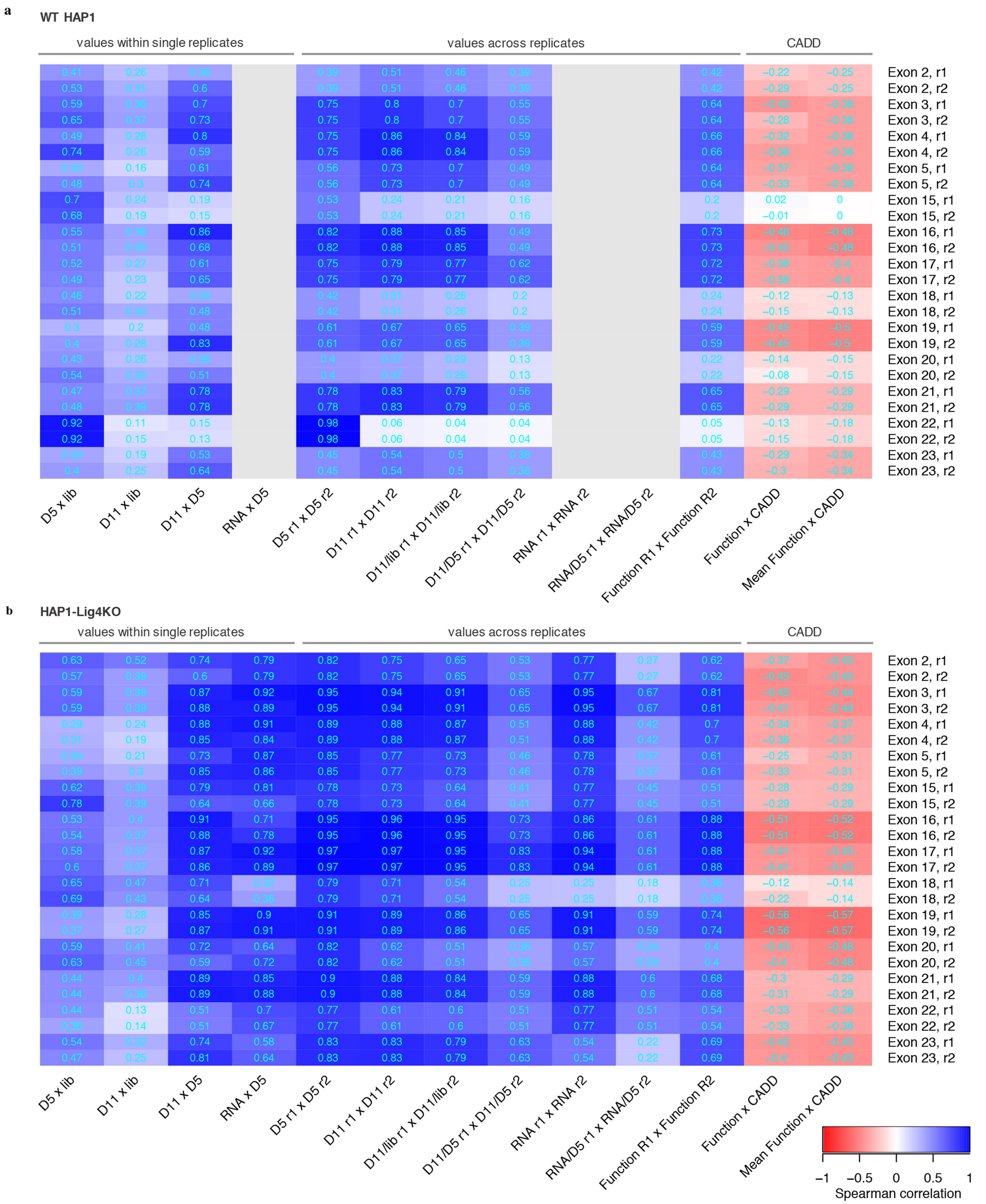
Extended Data Fig. 1 | CRISPR targeting of HDR pathway genes to confirm essentiality in HAP1 cells. a, Schematic, HAP1 cells are transfected with a plasmid expressing a gRNA and a Cas9-2A-puromycin cassette[24]. Owing to low transfection rates for HAP1 cells, puromycin selection reduces viable cells in all transfections. Over time, however, CRISPR targeting of non-essential genes leads to increased cell growth compared to CRISPR targeting of essential genes. **b**, HAP1 cell populations were transfected with a Cas9/gRNA plasmid either targeting the non-essential gene *HPRT1* (control) or exon 17 of *BRCA1* on day 0. Successfully transfected cells were selected with puromycin (days 1–4) and cultured until imaging on day 7, at which point cells were imaged. Images are representative of two transfection replicates. **c**, Cell viability of HAP1 cells transfected with Cas9/gRNA constructs targeting different HDR genes and controls (*HPRT1*, *TP53*) was measured using the CellTiterGlow assay. Luminescence is proportional to the number of living cells in each well when the assay is performed. Triplicate wells for each gRNA at each time point were processed, quantified on a plate reader and averaged. Error bars show the standard error of the mean. gRNA sequences are included in Supplementary Table 3. **d**, The targeted *BRCA1* exon 17 locus was deeply sequenced from a population of transfected cells sampled on day 5 and day 11. The fold-change from day 5 to day 11 for each editing outcome observed at a frequency over 0.001 in day 5 sequencing reads is plotted.



Extended Data Fig. 2 | Analysis of Cas9-induced indels observed in *BRCA1* SGE experiments. Variants observed in gDNA sequencing were included in this analysis if (i) they aligned to the reference with either a single insertion or deletion within 15 bp of the predicted Cas9 cleavage site and (ii) were observed at a frequency greater than 1 in 10,000 reads in both replicates. **a**, Histograms show the number of unique indels observed of each size, with negative sizes corresponding to deletions. More unique indels were observed in wild-type HAP1 cells compared to HAP1-LIG4^{KO} cells for exons compared (wild-type data for exon 22 was excluded). **b**, Day 11 over day 5 indel frequencies were normalized to the median synonymous SNV in each replicate and then averaged across replicates to measure selection on each indel. The distribution of selective effects is shown for each experiment as a histogram, in which indels are coloured by whether their size was divisible by 3 (that is, 'in-frame' versus 'frameshifting'). Whereas frameshifting variants were consistently depleted, some exons were tolerant to in-frame indels.

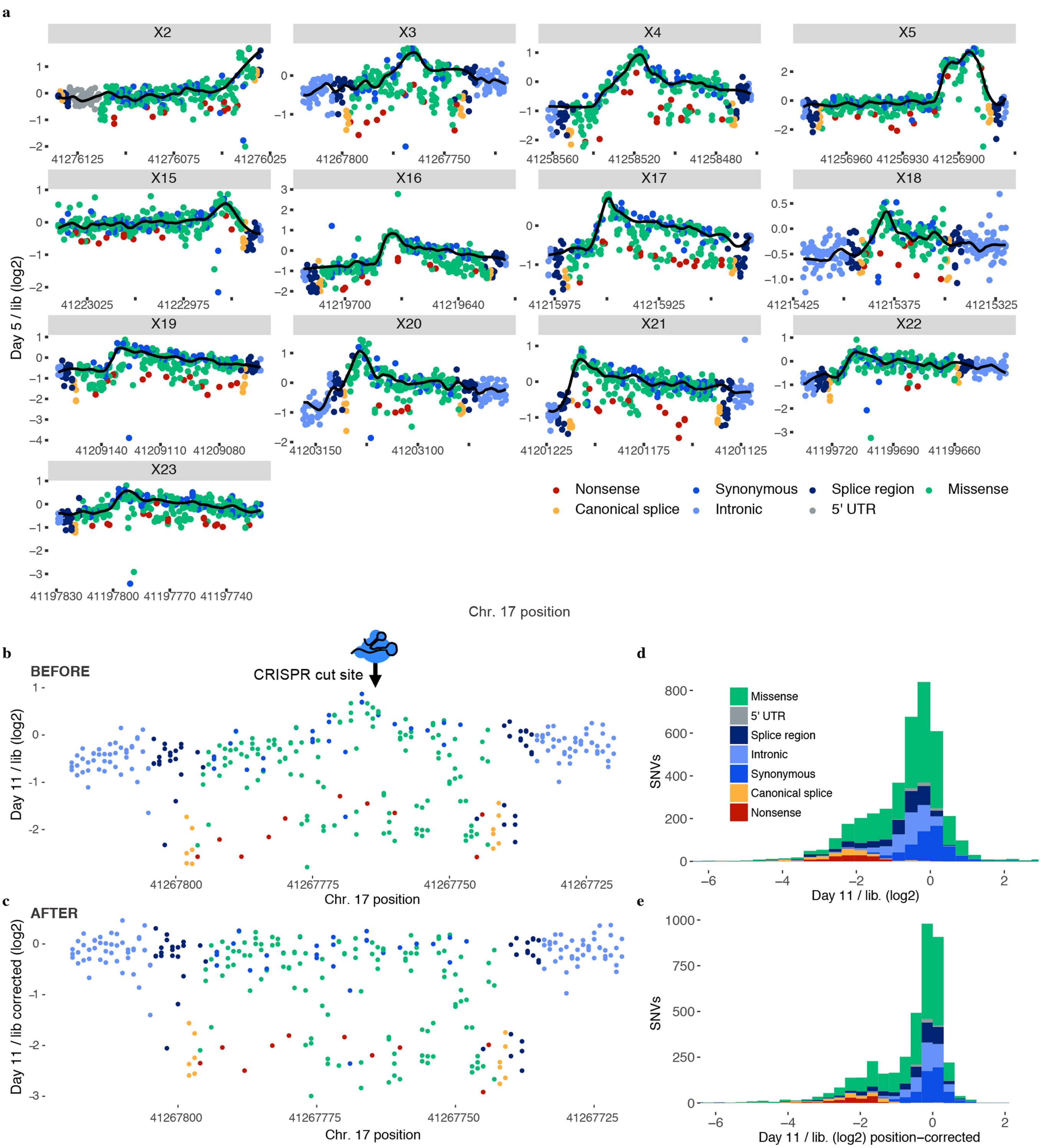


Extended Data Fig. 3 | HAP1 cell line optimizations for saturation genome editing to assay essential genes. a, A gRNA targeting Cas9 to the coding sequence of *LIG4*, a gene integral to the non-homologous end-joining pathway, was cloned into a vector co-expressing Cas9-2A-GFP[24]. Wild-type HAP1 cells were transfected, and single GFP-expressing cells were sorted into wells of a 96-well plate. Eight monoclonal lines were grown out over a period of three weeks and screened using Sanger sequencing for frameshifting indels in *LIG4*. The Sanger trace shows the frameshifting deletion present in the clonal line chosen for subsequent experiments, referred to as HAP1-LIG4^{KO}. **b**, To purify HAP1 cells for haploid cells, live cells were stained for DNA content with Hoechst 34580 and sorted using a gate to select cells with the lowest DNA content, corresponding to 1*n* cells in G1. **c**, The fraction of all possible SNVs scored is shown for each exon. SNVs were excluded mainly due to proximity to the HDR marker and/or poor sampling (Methods). **d**, **e**, Measurements across replicates are plotted for exon 17 SNVs assayed in HAP1-LIG4^{KO} cells to show correlations of day 5 frequencies (**d**) and day 11 over library ratios (**e**). **f**–**h**, Plots comparing SNV function scores across replicate experiments for exon 17 saturation genome editing experiments performed in unsorted wild-type HAP1 cells (**f**), HAP1-LIG4^{KO} cells (**g**), and wild-type HAP1 cells sorted on 1*n* ploidy (**h**). **i**, Function scores (averaged across replicates) are plotted to compare results for exon 17 experiments performed in wild-type 1*n*-sorted HAP1 cells and HAP1-LIG4^{KO} cells. The number of SNVs plotted and the Spearman correlation is displayed for each plot (**d**–**i**).

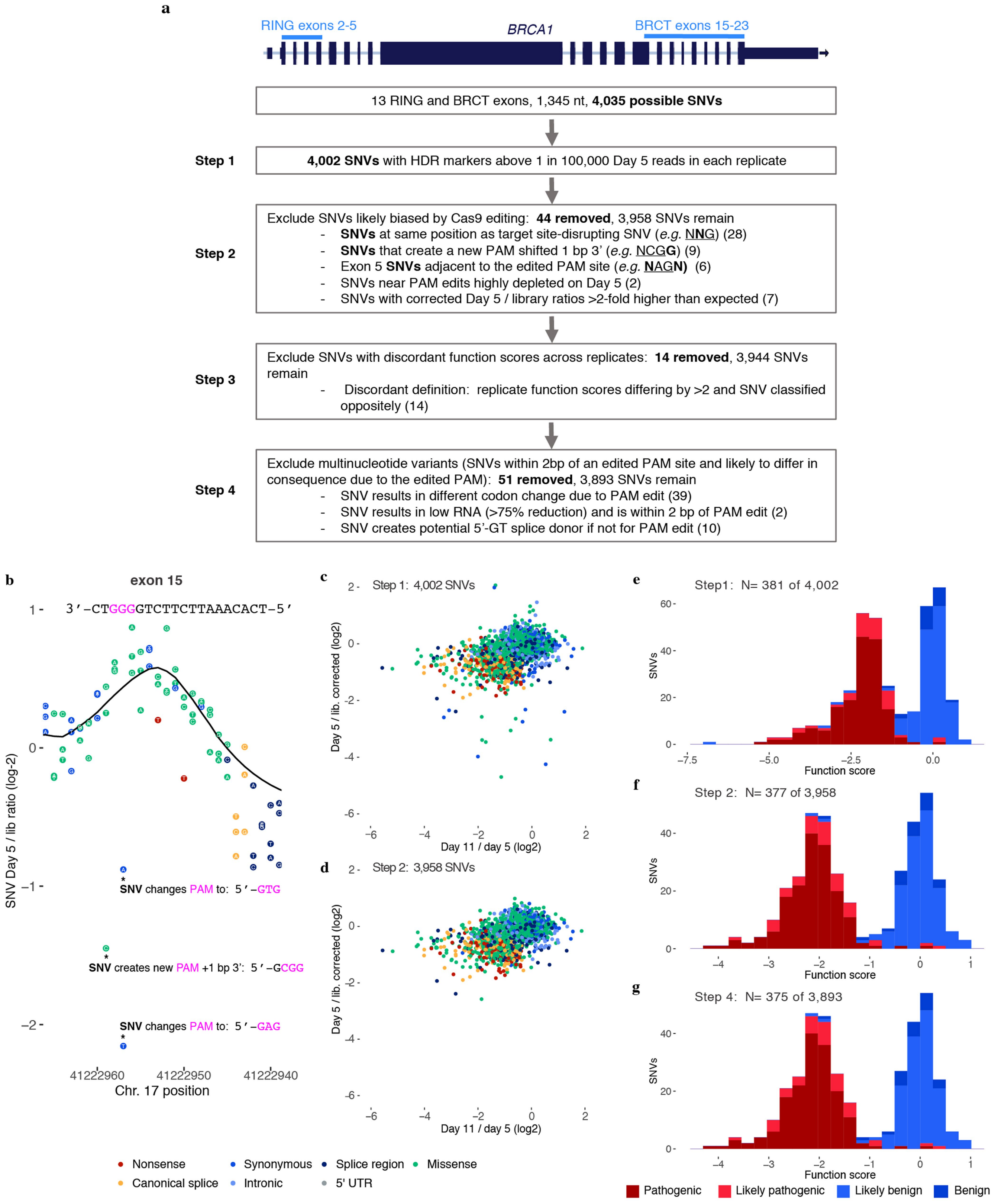


Extended Data Fig. 4 | See next page for caption.

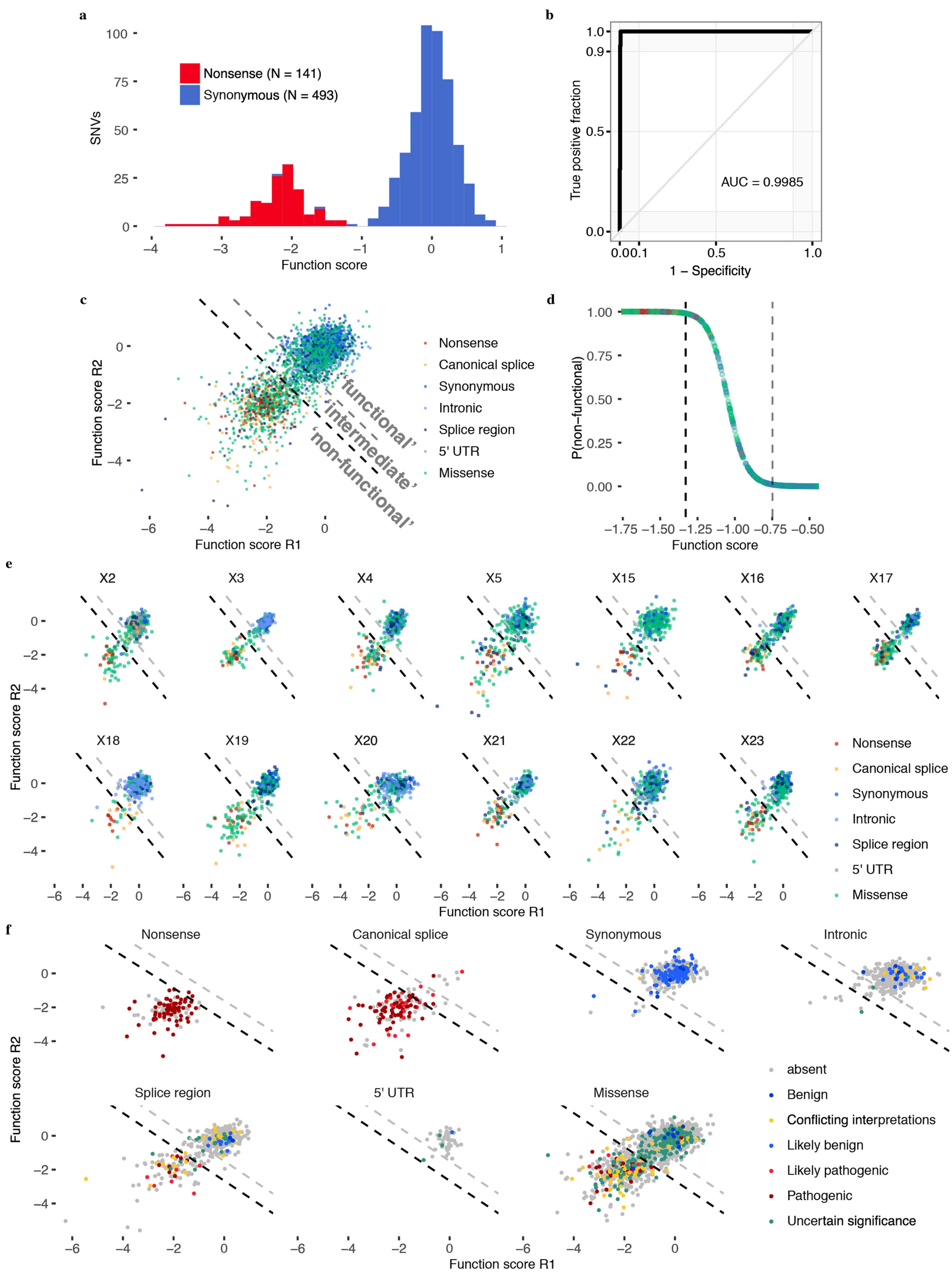
Extended Data Fig. 4 | Correlations for SNV measurements within single experiments, across transfection replicates, and to CADD scores for all SGE experiments. Heat maps indicate Spearman correlation coefficients for SNV measurements from experiments in wild-type HAP1 cells (**a**) and in HAP1-LIG4^{KO} cells (**b**). Grey boxes indicate absent RNA data from wild-type HAP1 cells. The four leftmost columns show how SNV frequencies correlate between samples from within a single replicate experiment. The unusually high correlations between exon 22 SNV frequencies in the plasmid library and in day 5 gDNA samples from wild-type HAP1 cells suggests plasmid contamination in gDNA. Indeed, primer homology to a repetitive element in the exon 22 library was identified. Consequently, the wild-type HAP1 exon 22 data was removed from analysis and a different primer specific to gDNA was used to prepare exon 22 sequencing amplicons from HAP1-LIG4^{KO} cells. The low HAP1-LIG4^{KO} correlations between exon 18 SNV frequencies in day 5 gDNA and RNA and between RNA replicates suggests RNA sample bottlenecking consequential to low RNA yields. Therefore, exon 18 RNA was also excluded from analysis. Consistent with the higher rates of HDR-mediated genome editing (Fig. 2a), replicate correlations (middle columns) were generally higher in HAP1-LIG4^{KO} cells than wild-type HAP1 cells. CADD scores predict the deleteriousness of each SNV, and are therefore negatively correlated with function scores (rightmost columns).



Extended Data Fig. 5 | Models of SNV editing rates across *BRCA1* exons to account for positional biases. Gene conversion tracts arising during HDR in human cells are short such that library SNVs are introduced to the genome more frequently near the CRISPR target site. We modelled this positional effect in our data for $n = 4,002$ SNVs (pre-filtering) using a LOESS regression fit on day 5 over library SNV ratios. **a**, Plots shown here are of the average of $n = 2$ replicates per exon, with the black line indicating the LOESS regression. By day 5, selective effects on gene function are evidenced by nonsense SNVs (red) appearing at lower frequencies compared to neighbouring SNVs. Therefore, to best approximate the SNV editing rate as a function of position alone (that is, the 'baseline'), the regression excluded SNVs that were selected against between day 11 and day 5 (see Methods). **b**, **c**, Day 11 over library SNV ratios were adjusted by the positional fit for each experiment in calculating function scores. This adjustment is illustrated here for an exon 3 replicate by plotting the day 11 over library ratio as a function of position before (**b**) and after (**c**) adjustment for ($n = 298$ SNVs). The elevated day 11 over library ratios for SNVs near the CRISPR cleavage site (indicated with an arrow) are corrected to achieve a more uniform baseline across the mutagenized region. **d**, **e**, The distributions of SNV day 11 over library ratios before and after accounting for positional effects are shown, coloured by mutational consequence ($n = 4,002$ SNVs, averaged across $n = 2$ replicates).

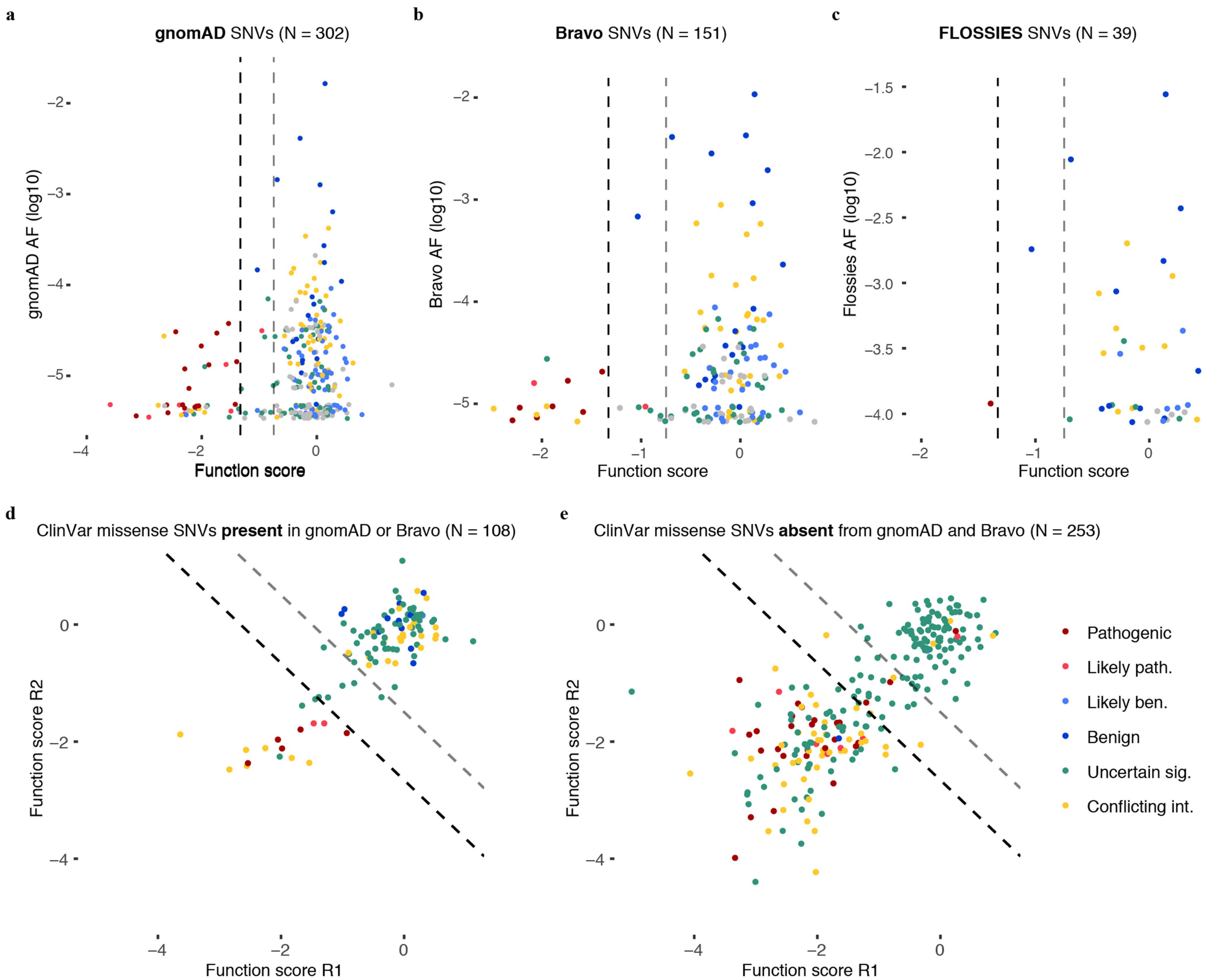


Extended Data Fig. 6 | SNV filtering to prevent erroneous functional classification. a, The flow chart describes filters used to produce the final SNV dataset and shows how many SNVs were removed at each step. **b**, Raw day 5 over library SNV ratios are shown for a portion of exon 15 to illustrate how re-editing biases necessitate filtering. The three depleted SNVs marked with asterisks create alternative PAM sequences that probably allow the Cas9–gRNA complex to re-cut the locus and cause their removal. For other SNVs, the fixed PAM edit (a GGG to GCG synonymous change) minimalizes re-editing. Alternative PAM sequences created by each indicated SNV are shown in magenta. The LOESS regression curve in shown in black. **c**, **d**, Plots show the relationship between day 5 over library and day 11 over day 5 ratios before (**c**) and after (**d**) filtering steps 1 and 2. Filtering removes outliers because editing biases primarily affect the day 5 over library ratio. **e**–**g**, Histograms show the distributions of function scores for SNVs deemed 'pathogenic' or 'benign' in ClinVar at different stages of filtering. Scores in **e** are derived before normalization across exons.

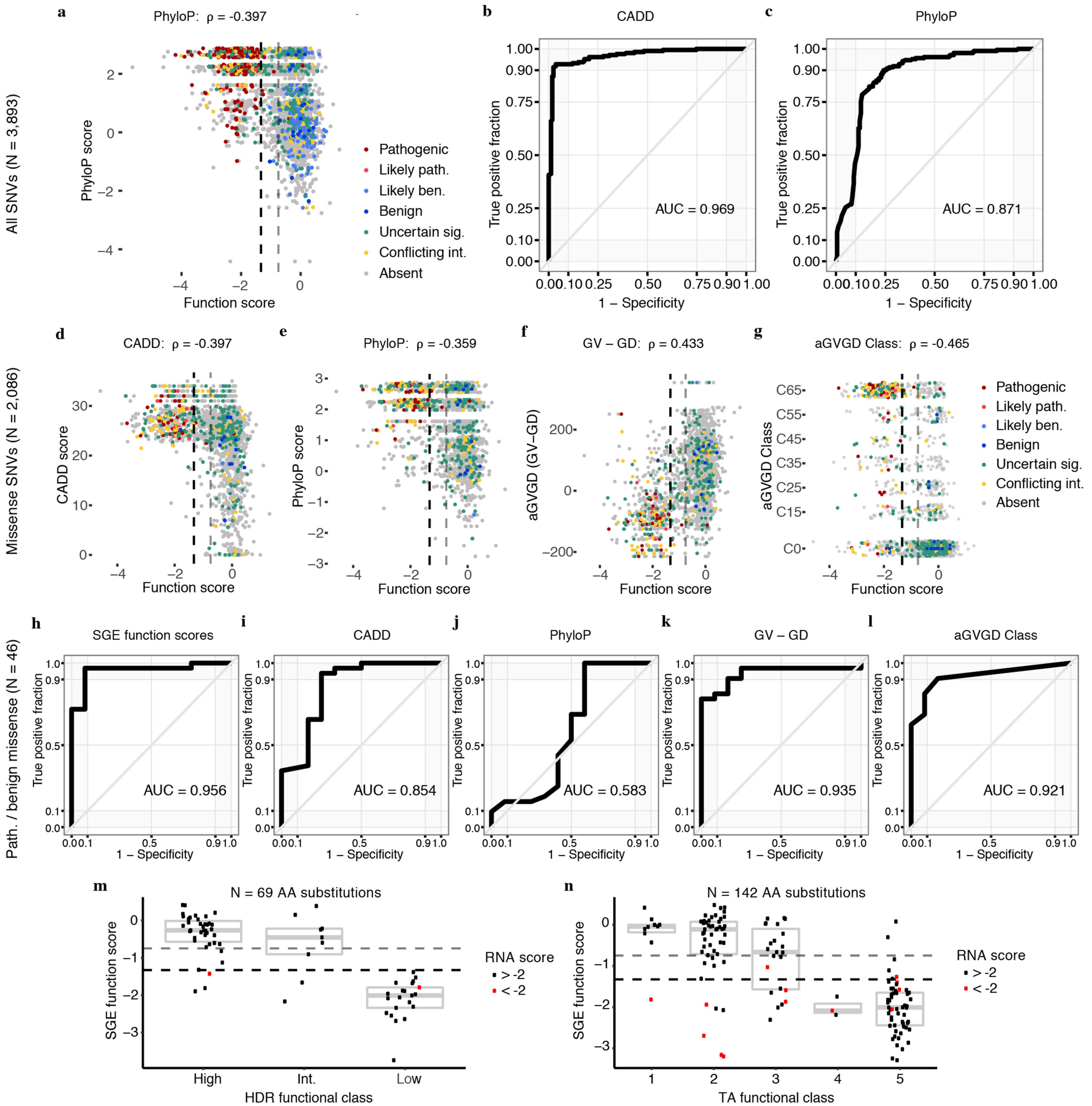


Extended Data Fig. 7 | See next page for caption.

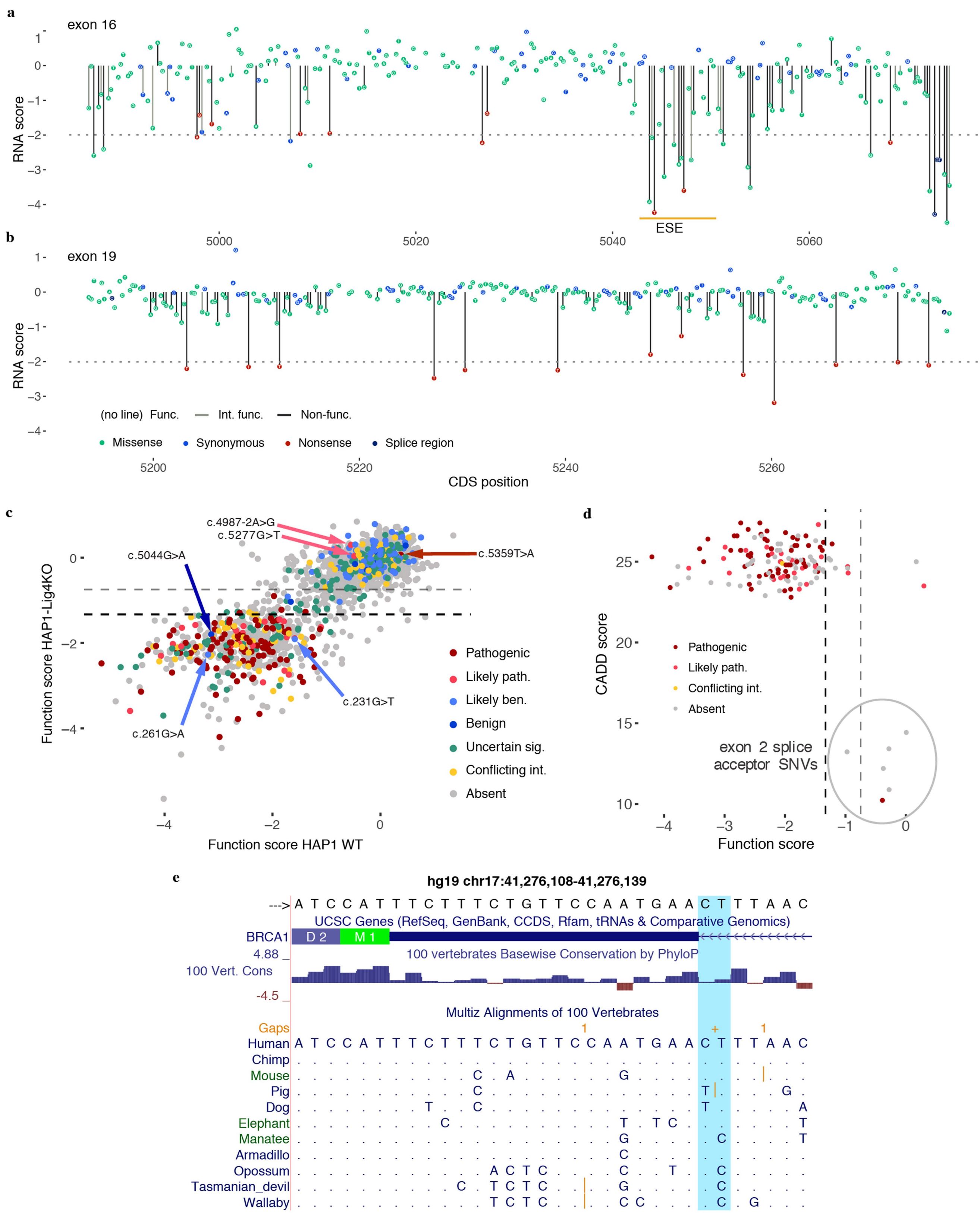
Extended Data Fig. 7 | Mixture modelling of scores to classify SNVs by functional effect. a, Distributions of 'non-functional' and 'functional' SNVs plotted here were defined respectively as all nonsense SNVs and all synonymous SNVs with RNA scores within 1 standard deviation of the median synonymous SNV. **b**, An ROC curve was generated using SGE function scores to distinguish the 634 'functional' and 'non-functional' SNVs defined in **a**. **c**, A two-component Gaussian mixture model was used to produce point estimates of the probability that each SNV was 'non-functional', P_{nf}, given its average function score across replicates. These *P* values are plotted in **d** against function scores for a subset of the data. Thresholds were set such that $P_{\mathrm{nf}} < 0.01$ corresponds to 'functional', and $P_{\mathrm{nf}} > 0.99$ corresponds to 'non-functional', and $0.01 < P_{\mathrm{nf}} < 0.99$ corresponds to 'intermediate' classification. Functional classification thresholds are drawn as dashed lines; black denotes the non-functional threshold and grey the intermediate threshold. **e**, **f**, SNV function scores across replicates are plotted for each exon with SNVs coloured by mutational consequence (**e**), and for each type of mutational consequence with SNVs coloured by ClinVar status (**f**). Using the optimal function score cutoff for all SNVs tested (Fig. 3b), sensitivities and specificities for distinguishing 'Pathogenic'/'Likely pathogenic' from 'Benign'/'Likely benign' ClinVar annotations for each type of mutation are as follows: 92.7% and 92.9% for missense SNVs ($n = 55$), 100% and 100% for splice region SNVs ($n = 23$), and 95.2% sensitivity for canonical splice site SNVs ($n = 83$; specificity not calculable).



Extended Data Fig. 8 | *BRCA1* SNVs observed more frequently in large-scale population sequencing are more likely to score as functional. **a–c**, SNV function scores are plotted against gnomAD (**a**), Bravo (**b**), and FLOSSIES (**c**) allele frequencies. **a**, Among the 302 SNVs assayed also present in gnomAD, higher allele frequencies associate with higher function scores (Wilcoxon signed-rank test, $P = 3.7 \times 10^{-12}$). **b**, Bravo is a collection of whole-genome sequences ascertained from 62,784 individuals through the NHLBI TOPMed program. Similarly to SNVs present in gnomAD, higher allele frequencies in Bravo correlate with higher function scores. **c**, FLOSSIES is a database of variants seen in targeted sequencing of breast cancer genes sampled from approximately 10,000 cancer-free women who are at least 70 years old. Only 1 of 39 assayed SNVs present in FLOSSIES scored as non-functional. **c**, **d**, Missense SNVs in ClinVar are separated by whether they have (**c**) or have not (**d**) been seen in either gnomAD or Bravo and function scores across replicates are plotted, with dashed lines demarcating functional classes. A higher proportion of ClinVar missense SNVs absent from gnomAD and Bravo score as non-functional (50.6% versus 15.7%; Fisher's exact test, $P = 1.80 \times 10^{-17}$).



Extended Data Fig. 9 | SGE function scores correlate with computational metrics and perform favourably at predicting ClinVar annotations. a, SNV function scores are plotted against mammalian phyloP scores, with colours indicative of ClinVar status (Spearman's correlation shown). **b**, **c**, ROC curves show the performance of CADD scores and phyloP scores for discriminating ClinVar 'pathogenic' and 'benign' SNVs (including 'likely'), as described in Fig. 3b for SGE data. **d**–**g** Plots as in **a**, but for missense SNVs only, showing correlations between SGE function scores and CADD[28] scores, phyloP scores[29], Grantham differences (Grantham amino acid variation minus Grantham amino acid deviation; GV − GD), and align-GVGD classifications[47]. Missense SNV function scores also correlate with SIFT scores[48] ($\rho = 0.363$) and PolyPhen-2 scores[49] ($\rho = -0.277$). (Spearman's correlation, $P < 1 \times 10^{-37}$ for all correlations). **h**–**l**, ROC curves assess the performance of SGE function scores and each indicated metric at distinguishing firmly 'pathogenic' and 'benign' missense SNVs (not including 'likely'). **m**, **n**, SGE scores for missense variants are plotted against results from homology-directed repair assays[9,31] (**m**) and results from transcriptional activation assays[12] (**n**). In cases where multiple SNVs assayed lead to same amino acid substitution, function scores were averaged and coloured red if either SNV had an RNA score less than −2. Box plots depict the sample median (line) and the interquartile range (box).



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Evidence supporting SNV scores in discordance with ClinVar classifications. **a**, **b**, Complete maps of RNA scores for exons 16 (**a**) and exon 19 (**b**) reveal highly variable sensitivity to RNA depletion. The location of the strongest predicted exonic splice enhancer in exon 16 is indicated by the orange line[36]. **c**, Function scores (means from two replicates) are plotted to compare results from preliminary experiments in wild-type HAP1 to those in HAP1-LIG4^{KO}. Data are shown only for experiments with Spearman's correlations between replicates greater than 0.50 in wild-type HAP1 cells ($n = 2{,}096$ SNVs; exons 3, 4, 5, 16, 17, 19, 21). Discordantly classified SNVs are indicated with arrows. c.19–2A>G was the only firmly discordant SNV for which the function score could not be corroborated in wild-type HAP1, consequent to low reproducibility of exon 2 wild-type function scores. Indeed, c.19–2A>G scored highly variably between wild-type replicates. **d**, The sequence-function map of exon 21 is shown with the function scores for the two 'pathogenic' SNVs observed in linkage indicated. Dashed lines demarcate functional classifications. **c**, Function scores are plotted against CADD scores for all canonical splice SNVs assayed, coloured by ClinVar status. The six possible exon 2 splice acceptor SNVs (circled) have the lowest CADD scores among all canonical splice SNVs assayed, and none score as 'non-functional'. **e**, A USCS Genome Browser shot shows the PhyloP conservation track and selected mammalian sequence alignments for the exon 2 acceptor region, with the canonical acceptor site nucleotides highlighted in light blue (hg19 chr17:41,276,108–41,276,139). Multiple mammalian species are identified that have a G at position c.19–2 of the human transcript (corresponding to a C in the plus-strand orientation shown).

ARTICLE

https://doi.org/10.1038/s41586-018-0552-x

Erythro–myeloid progenitors contribute endothelial cells to blood vessels

Alice Plein[1,3], Alessandro Fantin[1,3], Laura Denti[1], Jeffrey W. Pollard[2] & Christiana Ruhrberg[1]*

The earliest blood vessels in mammalian embryos are formed when endothelial cells differentiate from angioblasts and coalesce into tubular networks. Thereafter, the endothelium is thought to expand solely by proliferation of pre-existing endothelial cells. Here we show that a complementary source of endothelial cells is recruited into pre-existing vasculature after differentiation from the earliest precursors of erythrocytes, megakaryocytes and macrophages, the erythro-myeloid progenitors (EMPs) that are born in the yolk sac. A first wave of EMPs contributes endothelial cells to the yolk sac endothelium, and a second wave of EMPs colonizes the embryo and contributes endothelial cells to intraembryonic endothelium in multiple organs, where they persist into adulthood. By demonstrating that EMPs constitute a hitherto unrecognized source of endothelial cells, we reveal that embryonic blood vascular endothelium expands in a dual mechanism that involves both the proliferation of pre-existing endothelial cells and the incorporation of endothelial cells derived from haematopoietic precursors.

Blood vessels distribute oxygen, nutrients, hormones and immune cells through the vertebrate body and help to remove waste molecules. Accordingly, the formation of functional blood vessels during embryogenesis is a prerequisite for vertebrate life. Endothelial cells (ECs) form the inner lining of blood vessels; they first arise from mesenchymal precursors termed angioblasts on embryonic day (E)7.0 in mice[1]. After condensing into the yolk sac vasculature and paired dorsal aortae, ECs proliferate within existing endothelium to increase vascular diameter, sprout into avascular tissue areas or remodel into smaller vessels by intussusceptive growth[1]. The current consensus is therefore that embryonic ECs are a self-contained cell lineage that expands without contribution from new angioblasts or circulating precursors. By contrast, circulating endothelial progenitors have been proposed to exist in adult vertebrates, although their relationship to myeloid cells remains controversial[2].

In addition to their primary roles in the innate immune system, myeloid cells such as monocytes and macrophages also modulate vascular growth[3]. For example, the tissue-resident macrophages of the embryonic mouse brain, termed microglia, contact ECs at the tips of neighbouring vessel sprouts to promote their anastomosis into perfused vessel loops[4]. By contrast, no direct contribution of myeloid cells to embryonic vascular endothelium has been reported; thus, genetic lineage tracing with myeloid *Vav* or *Lyz2* (also known as *Lysm*) promoters does not mark embryonic blood vascular endothelium in mice[5,6].

Most tissue-resident macrophages arise from EMPs that form in the extra-embryonic yolk sac during embryogenesis and also serve as precursors for erythrocytes and megakaryocytes[7–11]. In mice, an early wave of EMPs, also referred to as primitive haematopoietic progenitors, buds from the yolk sac endothelium between E7.0 and E8.25 and differentiates by E9.0 without monocytic intermediates into yolk sac macrophages[7,10,12–14]. These macrophages colonize the embryo proper to generate tissue-resident macrophages, such as microglia in the brain or Langerhans cells in the epidermis[13]. A later wave of EMPs buds from the yolk sac endothelium from E8.25 onwards and colonizes the liver after the embryonic circulation has been established[7,8,11,14,15]. These later-born EMPs expand in the liver into monocytes that subsequently differentiate into tissue-resident macrophages in many organs except the brain[7,13].

Csf1r lineage tracing identifies an EC subset

To target early EMPs[7,10,12], microglia[16,17] and other differentiated myeloid cells[18], we and others have used a transgene that expresses CRE recombinase under the promoter for the myeloid lineage gene *Csf1r* (also known as *Fms*), which encodes the colony-stimulating factor 1 receptor CSF1R. Microglia appear as single YFP$^+$ cells in hindbrains from *Csf1r-iCre* mouse embryos carrying the *Rosa^{Yfp}* recombination reporter, with microglia and ECs also stained for isolectin B4 (IB4)[16] (Fig. 1a). As previously shown[4], the number of IB4$^+$YFP$^+$ microglia peaked in the hindbrain subventricular zone at E11.5, when vessels fuse into the subventricular vascular plexus (SVP) (Fig. 1b). Unexpectedly, we also observed sporadic, elongated IB4$^+$YFP$^+$ cells that appeared bound into the endothelium and increased steadily in number during SVP expansion (Fig. 1a–c; Extended Data Fig. 1a). *Csf1r-iCre* targeting of vessel-bound cells was not an artefact caused by spontaneous *Rosa^{Yfp}* recombination or unspecific immunostaining, because littermates that did not carry *Csf1r-iCre* lacked YFP staining (Fig. 1a). Furthermore, hindbrain imaging from mice carrying *Csf1r-iCre* with *CAG-Cat-Egfp* or *Rosa^{tdTom}* as alternative recombination reporters confirmed targeting of both microglia and vessel-bound elongated cells (Extended Data Fig. 1b, c). The tamoxifen-induced activation of CRE, expressed from an independently generated *Csf1r-Mer-iCre-Mer* transgene that targets myeloid cells[19], also targeted vessel-bound cells in addition to microglia (Fig. 1d). Corroborating the endothelial identity of *Csf1r-iCre*-targeted, elongated vessel-bound cells, these cells expressed the EC markers ERG and PECAM1, had a similar morphology to ECs targeted with the endothelium-specific *Cdh5-CreER^{T2}* transgene, formed junctions with neighbouring ECs via the endothelial cadherin CDH5 and lacked both myeloid and pericyte markers (Fig. 1e, Extended Data Fig. 1d–f).

Csf1r-iCre-mediated EC targeting was not explained by *Csf1r* expression in brain ECs, because hindbrain ECs, unlike microglia, lacked expression of a *Csf1r-Egfp* transgene that faithfully reports *Csf1r* promoter activity[20,21] and accordingly did not contain CSF1R protein

[1]UCL Institute of Ophthalmology, University College London, London, UK. [2]MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh, UK. [3]These authors contributed equally: Alice Plein, Alessandro Fantin. *e-mail: c.ruhrberg@ucl.ac.uk

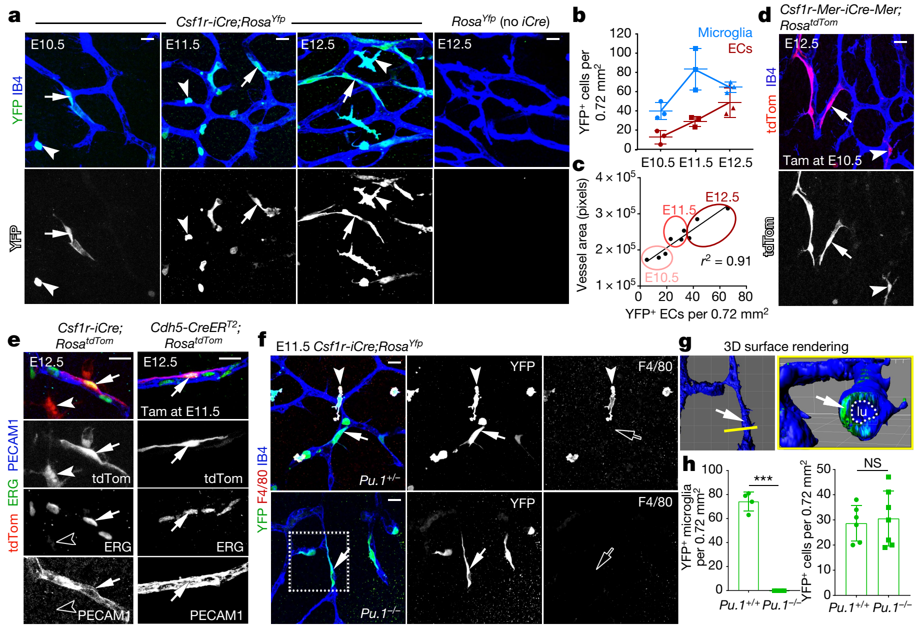


Fig. 1 | *Csf1r-iCre* lineage tracing identifies ECs in developing brain vasculature. a–c, Hindbrains of *Csf1r-iCre;Rosa*^{Yfp} mice at the indicated gestational stages. **a**, Whole-mount labelling for YFP and with IB4. **b**, Numbers of YFP$^+$IB4$^+$ single cells (microglia) and YFP$^+$IB4$^+$ vessel-bound cells (putative ECs) per 0.72 mm^2, mean ± s.d. **c**, Positive correlation between YFP$^+$ putative EC number and vessel area (r^2, coefficient of determination; goodness of fit, $P < 0.01$); each data point represents one hindbrain, $n = 3$ hindbrains for each group. **d**, **e**, Hindbrains from E12.5 embryos of the indicated genotypes, whole-mount labelled with the indicated markers and shown including tdTomato (tdTom) fluorescence; *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} (**d**) was tamoxifen-induced on E10.5 and *Cdh5-CreER*^{T2}*;Rosa*^{tdTom} (**e**) on E11.5; $n = 3$ hindbrains for each genotype. **f–h**, Hindbrains from *Csf1r-iCre;Rosa*^{Yfp} embryos on a *Pu.1*$^{+/+}$ versus *Pu.1*$^{-/-}$ background at E11.5, labelled for YFP and F4/80 together with IB4. The boxed area in **f** was 3D surface rendered and is shown in **g** en face and as a lateral view starting at the plane indicated by the yellow line; the vascular lumen (lu) is outlined. **h**, YFP$^+$ microglia (*Pu.1*$^{+/+}$, $n = 4$; *Pu.1*$^{-/-}$, $n = 3$) and ECs (*Pu.1*$^{+/+}$, $n = 6$; *Pu.1*$^{-/-}$, $n = 7$), mean ± s.d.; each data point represents one hindbrain; NS, not significant; ***$P < 0.0001$ (two-tailed unpaired *t*-test). Arrowheads, microglia; arrows, ECs. Solid and clear symbols indicate the presence or absence of marker expression, respectively. Scale bars: 20 μm (**a**, **d**, **f**), 50 μm (**e**).

(Extended Data Fig. 2a, b). Moreover, our analysis of published transcriptomic datasets[22] showed that *Csf1r* is not expressed in ECs from embryonic brain, liver or lung, whilst quantitative PCR with reverse transcription (RT–qPCR) analysis of tdTomato$^+$ ECs isolated by fluorescence-activated cell sorting (FACS) confirmed that they expressed *Cdh5*, but not *Csf1r* or the myeloid gene *Spi1*, which encodes the PU.1 transcription factor (Extended Data Fig. 2c–g). The lack of endothelial *Csf1r* expression suggests that *Csf1r-iCre*-targeted brain ECs arise from precursors in which *Csf1r* is activated before their incorporation into hindbrain vasculature. These precursors cannot be differentiated myeloid cells such as microglia, whose formation is PU.1-dependent[4], because PU.1 deficiency did not reduce the number of *Csf1r-iCre*-targeted ECs in the hindbrain at E11.5 (Fig. 1f–h) or the striatum at postnatal day (P)0 (Extended Data Fig. 2h). We therefore investigated whether *Csf1r-iCre*-targeted ECs are derived from PU.1-independent, *Csf1r*-expressing precursors.

Csf1r lineage-traced ECs derive from EMPs

As EMPs are PU.1-independent[23], we investigated whether the formation of *Csf1r-iCre*-targeted ECs was spatiotemporally linked to the emergence of *Csf1r*-expressing EMPs from yolk sac haemogenic endothelium, which was visualized by staining for the EC marker VEGFR2. At E8.5, yolk sacs from *Csf1r-Egfp* embryos contained clusters of round EGFP$^+$VEGFR2$^+$ cells that protruded from the endothelium into the vascular lumen (Fig. 2a), consistent with previous work showing that FACS-isolated EMPs express both *Csf1r*[7] and *Vegfr2*[9], and that EMPs bud from the yolk sac endothelium[11]. *Csf1r-iCre* lineage tracing in yolk sacs at E8.5 similarly identified round YFP$^+$ cells that protruded into the vascular lumen, expressed VEGFR2, persisted in PU.1-deficient yolk sacs and expressed the EMP marker KIT[7] (Fig. 2b, c; Extended Data Fig. 3a, b). Even though EGFP expression could not be detected in *Csf1r-Egfp* yolk sac endothelium (Fig. 2a), *Csf1r-iCre;Rosa*^{Yfp} also targeted a subset of yolk sac ECs that lacked obvious

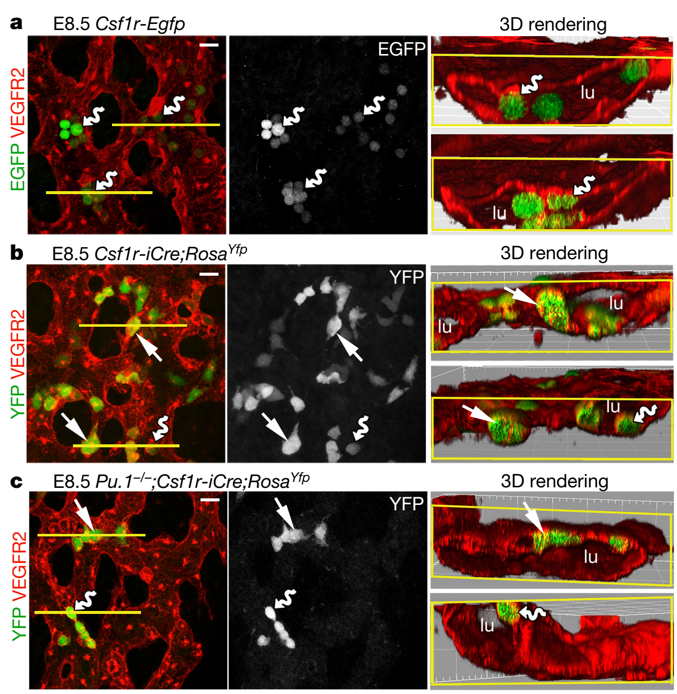


Fig. 2 | *Csf1r-iCre*-targeted ECs emerge concomitantly with EMPs in the yolk sac. E8.5 yolk sacs were whole-mount labelled with the indicated markers. **a**, *Csf1r-Egfp* yolk sacs. **b**, **c**, *Csf1r-iCre;Rosa*^{Yfp} yolk sacs on a *Pu.1*$^{+/+}$ (**b**) versus *Pu.1*$^{-/-}$ background (**c**). $n = 4$ yolk sacs for each genotype. The yellow lines mark the start of 3D-rendered lateral views. Wavy arrows indicate VEGFR2$^+$EGFP$^+$ and VEGFR2$^+$YFP$^+$ round EMPs or myeloid progenitors protruding from the vascular wall into the lumen (lu); straight arrows indicate YFP$^+$VEGFR2$^+$ flat cells within the vascular wall. Scale bars, 20 μm.

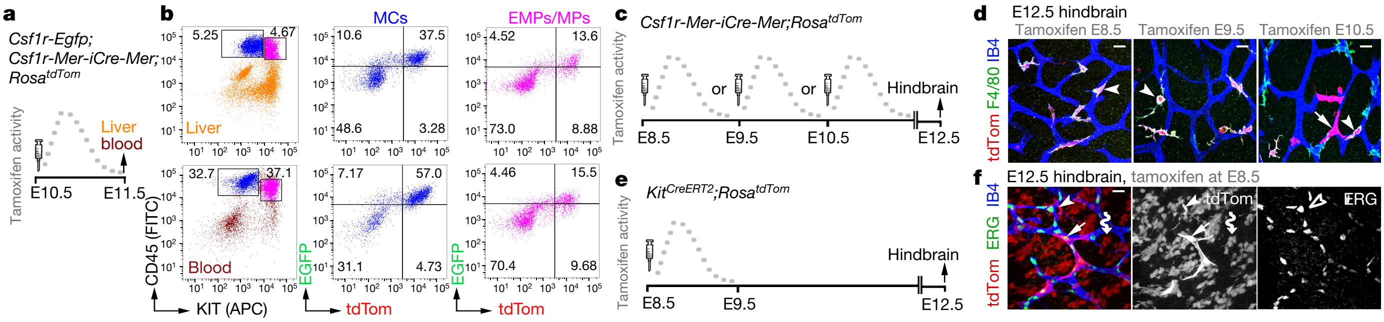


Fig. 3 | *Csf1r-iCre*-targeted hindbrain ECs emerge from intraembryonic EMPs. **a**, **b**, A pregnant *Csf1r-Egfp;Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} dam was injected with a single dose of tamoxifen on E10.5 (**a**) before FACS of E11.5 liver and blood cells (**b**) to gate the CD45^{high}KIT^{-} differentiated myeloid cell (MCs; blue) and CD45^{low}KIT^{+} EMP/myeloid progenitor (MPs; pink) populations for EGFP and tdTomato ($n = 4$ embryos). **c**–**f**, Pregnant *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} (**c**, **d**) and *Kit*^{CreERT2}*;Rosa*^{tdTom} (**e**, **f**) dams were injected with a single dose of tamoxifen on the indicated days before whole-mount staining of E12.5 hindbrains for the indicated markers and imaging including tdTomato fluorescence; $n = 3$ independent experiments each. Arrows, tdTomato^{+} ECs; arrowheads, microglia; the clear arrowhead indicates lack of ERG expression in microglia; wavy arrows, cluster of tdTomato^{+}ERG^{-}IB4^{-} neural cells derived from *Kit*$^{+}$ neural progenitors. Scale bars, 20 μm.

KIT expression (Extended Data Fig. 3b), showing that they were not haemogenic ECs[24]. Similar to *Csf1r-iCre*-targeted hindbrain ECs, the lineage-traced yolk sac ECs were PU.1-independent (Fig. 2b, c).

The finding that EMP formation correlates with the emergence of *Csf1r-iCre*-targeted yolk sac ECs was corroborated by temporally restricted *Csf1r-Mer-iCre-Mer*-mediated lineage tracing. As tamoxifen-induced, CRE-mediated reporter recombination peaks approximately 6 h and ends 24 h after tamoxifen injection[25], we activated *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} in discrete temporal windows by single injections at E8.5, E9.5 or E10.5 before identifying lineage-traced cells in E12.5 yolk sacs (Extended Data Fig. 3c). Induction at all three stages labelled yolk sac macrophages (Extended Data Fig. 3d), consistent with their origin from *Csf1r*-expressing EMPs[7] and their maintenance of *Csf1r* expression[10,12]. In addition, induction at E8.5 or E9.5 yielded tdTomato^{+} ECs, whereas induction at E10.5 did not (Extended Data Fig. 3d). As EMPs are present in the yolk sac at E8.5 and E9.5, but move to the liver thereafter[13], their local availability makes them plausible precursors of *Csf1r-iCre*-labelled yolk sac ECs. Consistent with this, tamoxifen induction of a *Kit*^{CreERT2} knock-in allele at E8.5, when KIT^{+} early EMPs are still present in the yolk sac[7], lineage-traced both yolk sac ECs and macrophages (Extended Data Fig. 3e, f).

In contrast to early wave EMPs that remain in the yolk sac, the late wave EMPs that populate the embryo are reported to lack *Csf1r* expression, at least when they form in the yolk sac[7]. We therefore investigated whether late wave EMPs begin to express *Csf1r* after homing to the liver and whether they are the precursors of the *Csf1r-iCre*-targeted ECs that appear in the hindbrain from E10.5 onwards. Thus, we combined the *Csf1r-Egfp* expression reporter with *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} and

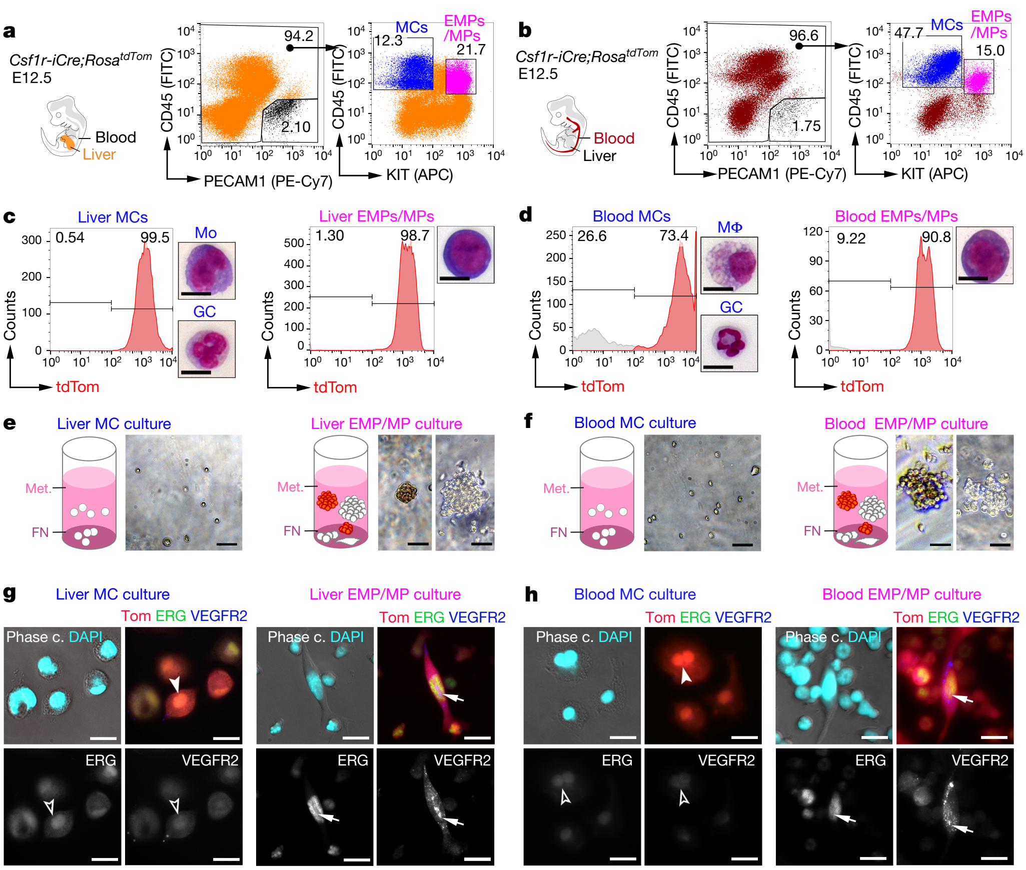


Fig. 4 | EMPs in the liver and blood give rise to ECs in vitro. **a**, **b**, FACS strategy to separate the differentiated myeloid cell and EMP/MP populations from E12.5 *Csf1r-iCre;Rosa*^{tdTom} liver (**a**) and blood (**b**) using antibodies for CD45 and KIT after excluding PECAM1^{+} cells to prevent EC contamination. **c**, **d**, tdTomato^{+} proportions in the FACS-isolated myeloid cell versus EMP/MP populations from liver and blood shown in **a**, **b** and Giemsa–Wright staining of representative cells. Mo, monocyte; GC, granulocyte; MΦ, macrophage. **e**–**h**, Bright-field images of myeloid (white) and erythroid (rust-coloured) colonies (**e**, **f**) and immunofluorescence of adherent cells (**g**, **h**) after three days in methocult (met.) on fibronectin. Adherent cells were immunolabelled for ERG and VEGFR2, counterstained with DAPI and are shown together with tdTomato fluorescence (Tom). Arrows, tdTomato^{+} ECs; arrowheads, tdTomato^{+} myeloid cells; solid and clear symbols indicate high and low marker expression, respectively. Scale bars, 20 μm. $n = 3$ independent experiments.

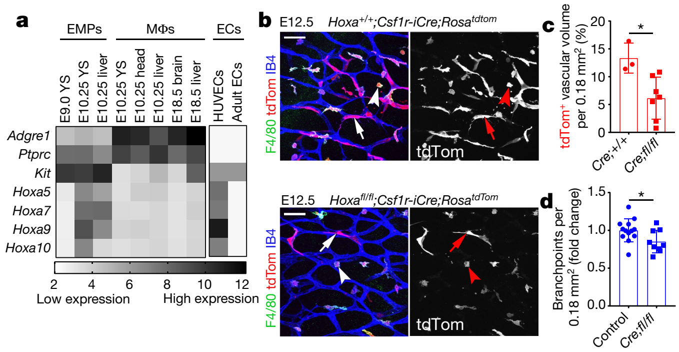


Fig. 5 | *Csf1r-iCre*-targeted ECs form in a *Hoxa*-dependent mechanism and promote vascularization of the embryonic hindbrain.
a, Transcriptomic analysis of the indicated cell populations for the indicated genes, based on published RNA-seq[9] and microarray data[30], shows that *Hoxa* transcripts are enriched in intraembryonic EMPs and perinatal ECs, respectively; white and black represent low and high relative gene expression, respectively; Mϕs, macrophages; YS, yolk sac; HUVECs, human umbilical vein ECs; *Adgre1* and *Ptprc* encode F4/80 and CD45, respectively; RNA-seq $n = 2$, except for E10.25 YS (4) and head Mϕs (3); microarray $n = 3$. **b**–**d**, E12.5 littermate hindbrains of the indicated genotypes. **b**, Whole-mount labelling for the indicated markers; RFP staining to visualize tdTomato shows fewer *Csf1r-iCre*-targeted ECs in mutant compared to control hindbrains; arrows, tdTomato$^+$ ECs; arrowheads, microglia. Scale bars, 50 μm. **c**, tdTomato$^+$ relative to IB4$^+$ EC volume in *Hoxa*$^{+/+}$ ($n = 3$) versus *Hoxa*$^{fl/fl}$ ($n = 7$) hindbrains on a *Csf1r-iCre;Rosa*^{tdTom} background, mean ± s.d. **d**, SVP complexity, measured as fold change in vascular branchpoints in *Hoxa*$^{fl/fl}$;*Csf1r-iCre* ($n = 9$) relative to control hindbrains (pooled *Csf1r-iCre*$^+$;*Hoxa*$^{+/+}$ and *Csf1r-iCre*$^-$ of any *Hoxa* genotype, $n = 13$), mean ± s.d. Each data point represents one hindbrain; $*P = 0.0184$ (**c**), $*P = 0.0323$ (**d**) (two-tailed unpaired *t*-test).

induced CRE-mediated recombination at E10.5; 24 h later, we used FACS to separate the differentiated myeloid cells from EMPs and EMP-derived myeloid progenitors[7] contained in the liver or blood (Fig. 3a, b). The differentiated myeloid cell populations from both sources contained tdTomato$^+$EGFP$^+$ cells, as expected, but tdTomato$^+$EGFP$^+$ cells were also present in the EMP/myeloid progenitor populations from both liver and blood (Fig. 3a, b; Extended Data Fig. 3g–i). These findings suggest that a subset of intraembryonic EMPs expresses *Csf1r* and can access organs such as the hindbrain via the circulation.

To determine whether the intraembryonic presence of *Csf1r*-expressing late wave EMPs correlated with the emergence of *Csf1r-iCre*-targeted hindbrain ECs, we visualized tdTomato expression in hindbrains from E12.5 *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} mice after tamoxifen induction at E8.5, E9.5 or E10.5 (Fig. 3c). The hindbrain vasculature contained tdTomato$^+$ ECs following induction at E10.5, but not at E8.5 or E9.5, even though the *Csf1r*-expressing microglia were targeted at all stages (Fig. 3d). *Kit*^{CreERT2} induction at E8.5 also caused microglia targeting (Fig. 3e, f), consistent with microglia arising from yolk sac macrophages generated at around E8.5 from KIT$^+$ early wave EMPs[7]. Induction of *Kit*^{CreERT2} at E8.5, when late wave EMPs begin to arise in the yolk sac[7], also yielded tdTomato$^+$ ECs in the hindbrain at E12.5 (Fig. 3e, f), confirming that yolk sac-born EMPs can give rise to intraembryonic ECs. Lineage tracing from three independent *Cre* alleles therefore suggests that EMPs give rise to both yolk sac and hindbrain ECs.

Csf1r-expressing EMPs give rise to ECs in vitro

The myeloid and erythroid potential of EMPs has been demonstrated through in vitro differentiation[11,26]. Using similar assays, we compared the endothelial potential of FACS-isolated differentiated myeloid cell and EMP/myeloid progenitor populations from E12.5 *Csf1r-iCre;Rosa*^{tdTom} liver and blood, while ensuring that we were excluding contamination by PECAM1$^+$ ECs (Fig. 4a, b). Both cell populations were mostly tdTomato$^+$ (Fig. 4c, d). As expected[7], the EMP-containing population comprised round cells with a large nucleus and little cytoplasm, whereas the myeloid cell population contained granulocytes, in addition to monocytes in the liver and macrophages in the blood (Fig. 4c, d). For cell culture, we used methocult to promote the formation of haematopoietic colonies, but included a fibronectin substrate to facilitate EC differentiation. Differentiated myeloid cells persisted in these cultures as single round or amoeboid cells (Fig. 4e, f) that were tdTomato$^+$ERG^{low}VEGFR2^{low} (Fig. 4g, h; antibody controls in Extended Data Fig. 4a, b). By contrast, both liver and blood EMPs formed myeloid and erythroid cell colonies in suspension (Fig. 4e, f) and additionally gave rise to single adherent cells that appeared spindle-shaped, were tdTomato$^+$ERG^{high}VEGFR2^{high} and lacked myeloid cell markers, consistent with an EC identity (Fig. 4g, h; Extended Data Fig. 4c). Together, these experiments demonstrate that EMPs have endothelial potential alongside their known haematopoietic capacity.

Csf1r lineage ECs support blood vessel growth

Hoxa cluster genes modulate haematopoiesis[27] and are upregulated in perinatal ECs compared to adult ECs[28] (Fig. 5a); HOXA9 also promotes EC differentiation from progenitors in ischaemic disease in adults[29]. Our analysis of published transcriptomic data[9,30] revealed that *Hoxa* transcripts are enriched in E10.25 EMPs compared to E9.0 EMPs and macrophages (Fig. 5a). To investigate whether *Hoxa* deficiency impairs the formation of EMP-derived hindbrain ECs, we combined *Csf1r-iCre* with a conditional null *Hoxa* cluster mutation (*Hoxa*^{fl}) (Extended Data Fig. 5a). Gene copy analysis showed effective gene targeting in KIT$^+$ cells from *Csf1r-iCre;Hoxa*$^{fl/fl}$ mutants at E12.5 compared to control livers, but the number of liver CD45$^+$ cells, including differentiated myeloid cells, was not reduced (Extended Data Fig. 5b–f). *Hoxa* genes are therefore dispensable for myeloid cell specification from late wave EMPs. By contrast, fewer tdTomato$^+$ ECs, also derived from late wave EMPs, had formed in *Rosa*^{tdTom}-carrying *Csf1r-iCre;Hoxa*$^{fl/fl}$ mutant hindbrains compared to control hindbrains; moreover, SVP complexity was reduced in mutant hindbrains (Fig. 5b–d). Although we observed 20% fewer microglia in mutant hindbrains than in control hindbrains (Extended Data Fig. 5g–i), this is not likely to have contributed to the vascular defect, because even a 50% microglia reduction in *Csf1*$^{op/+}$ mutants did not reduce SVP complexity (Extended Data Fig. 5j–l). Together, these findings suggest that *Hoxa* cluster genes promote the formation of EMP-derived brain ECs, which in turn support normal brain vascular development.

Transcriptional signature of *Csf1r* lineage ECs

Csf1r-iCre-targeted ECs not only appeared morphologically similar to neighbouring ECs (Fig. 1), but also had similarly slow proliferation and overall cell cycle kinetics (Extended Data Fig. 6). Moreover, RNA sequencing (RNA-seq) analysis of FACS-isolated tdTomato$^+$ and tdTomato$^-$ ECs from E12.5 *Csf1r-iCre;Rosa*^{tdTom} embryos showed that they had largely similar transcriptomes, with only a few differentially expressed genes, including the expected difference in the *tdTomato* transcript (Fig. 6a–c; Extended Data Fig. 7a). Corroborating their endothelial identity, tdTomato$^+$ ECs lacked markers for differentiated myeloid cells and other non-EC lineages, but expressed core EC transcripts at similar levels to tdTomato$^-$ ECs (Fig. 6d, e). Amongst the differentially expressed genes, markers typical of EC specialization, such as ephrins and EPH receptors regulating arteriovenous differentiation, were under-represented in tdTomato$^+$ ECs (Fig. 6e). This observation is consistent with *Csf1r-iCre*-targeted ECs being derived from progenitors that are recruited into preformed vascular endothelium. Whereas brain EC markers (for example, *Slc2a1*) were under-represented in the embryo-wide tdTomato$^+$ EC population, liver EC markers (for example, *Oit3*, *Mrc1*) were over-represented, including early markers of liver sinusoidal differentiation (*Stab2*, *Lyve1*)[31] (Fig. 6c, f; Extended Data Fig. 7b, c). Similar expression of *Oit3* and *Mrc1* in tdTomato$^+$ and tdTomato$^-$ liver ECs (Extended Data Fig. 7d) suggests that the over-representation of liver EC transcripts in the total embryonic tdTomato$^+$ EC population reflects their preferential contribution to

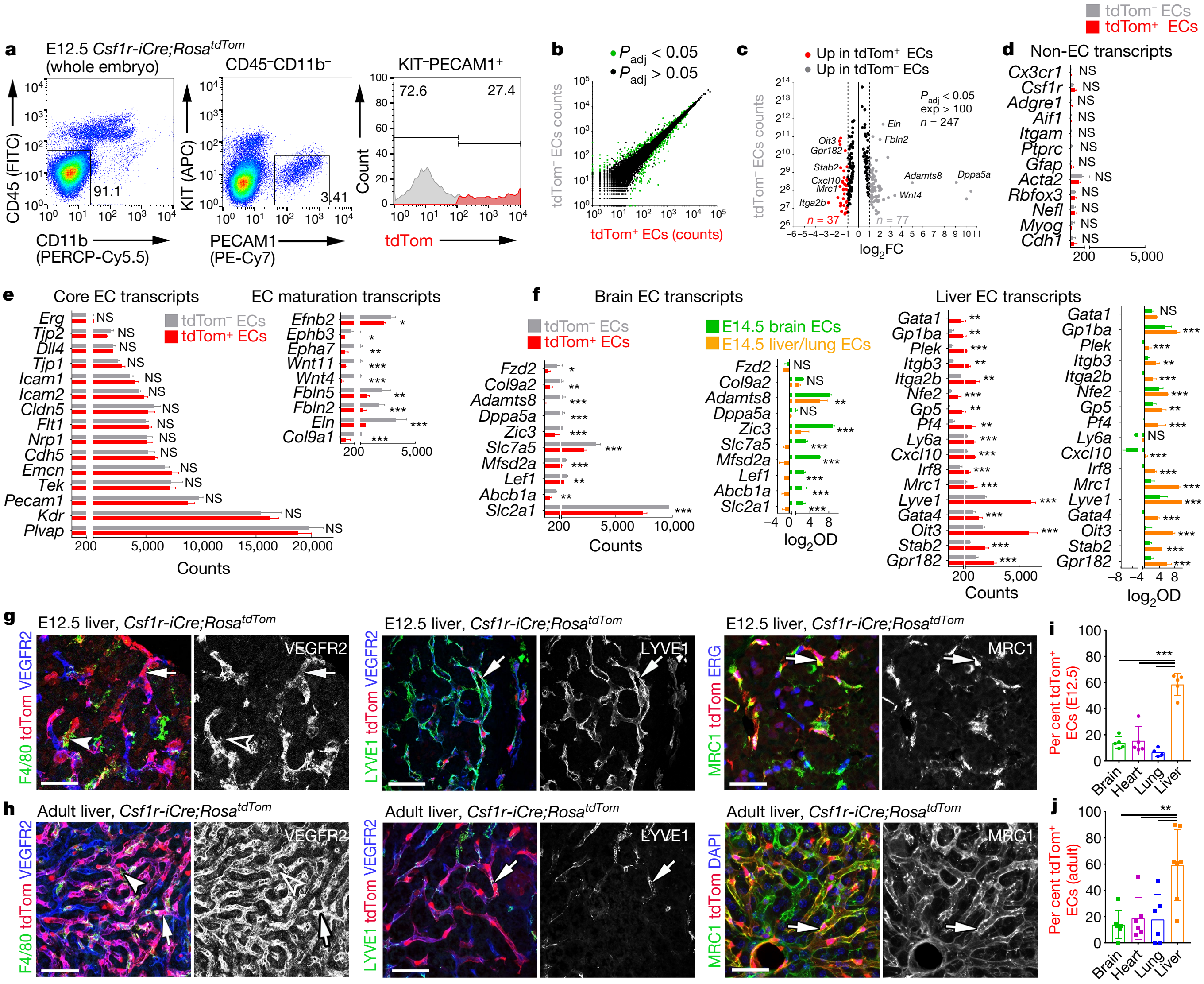


Fig. 6 | The *Csf1r-iCre*-targeted EC population has a core endothelial transcription signature with an increase in liver EC transcripts and persists in adult organs. a–f, Transcriptomic analysis. **a**, FACS strategy to isolate tdTomato⁻ and tdTomato⁺ ECs from E12.5 *Csf1r-iCre;Rosa^{tdTom}* embryos for RNA-seq. **b**, Graphic representation of genes for which expression was significantly different (green dots) or similar (black dots) between both EC populations. **c**, Volcano plot of significantly differentially expressed transcripts with more than 100 counts per transcript; selected genes are named; grey and red data points represent transcripts in tdTomato⁻ ECs with at least twofold over- or under-representation, respectively. **d–f**, Relative expression levels for markers typical of myeloid (*Cx3cr1*–*Ptprc*), astrocytic (*Gfap*), smooth muscle (*Acta2*), neuronal (*Rbfox3*, *Nefl*), skeletal muscle (*Myog*) or epithelial (*Cdh1*) differentiation (**d**), for EC core and maturation markers (**e**) and for representative brain and liver EC specialization markers compared to brain versus liver/lung ECs microarrays[30] (**f**); mean ± s.d.; RNA-seq, $n = 3$ embryos (DESeq2; Benjamini–Hochberg's multiple comparisons test for P value adjustment, P_{adj}); microarray, $n = 5$ organs (two-way ANOVA, Bonferroni's multiple comparisons test); NS, not significant, $*P < 0.05$, $**P < 0.01$, $***P < 0.0001$; see Source Data for exact values. **g**, **h**, *Csf1r-iCre;Rosa^{tdTom}* E12.5 (**g**) and adult (**h**) liver cryosections, labelled for the indicated markers and RFP to visualize tdTomato, including DAPI counterstaining in **h**; $n = 3$ independent experiments. Arrows, tdTomato⁺ ECs; arrowheads, macrophages; clear arrowheads indicate that macrophages lack VEGFR2. Scale bar, 50 μm. **i**, **j**, FACS of *Csf1r-iCre;Rosa^{tdTom}* E12.5 (**i**) and adult (**j**) brain, heart, lung and liver to determine their relative tdTomato⁺ EC contributions; mean ± s.d.; $n = 5$ organs each (**i**; except lung, $n = 4$), $n = 6$ organs each (**j**; except liver, $n = 7$); each data point represents one organ; $***P < 0.0001$ (**i**); $**P = 0.0023$, 0.0066, 0.00541 (**j**) for liver versus brain, heart, lung, respectively (one-way ANOVA with Tukey's multiple comparisons test).

liver vasculature. Immunostaining and FACS of *Csf1r-iCre;Rosa^{tdTom}* E12.5 and E18.5 embryos confirmed that tdTomato⁺ ECs were more prevalent than tdTomato⁻ ECs in liver endothelium (Fig. 6g, i; Extended Data Figs. 8, 9a, b). As liver EC specialization markers were present in both tdTomato⁻ and tdTomato⁺ liver ECs at E12.5 (Fig. 6g; Extended Data Fig. 8a), liver ECs from two distinct origins appear to undergo similar organ-specific EC differentiation.

Csf1r lineage ECs persist in multiple adult organs

Immunostaining and FACS analyses at E12.5 and E18.5 showed that *Csf1r-iCre*-targeted ECs were also present in the heart and lung vasculature at similar levels to the brain (Fig. 6i; Extended Data Figs. 8, 9a, b). Corresponding immunostaining and FACS analyses showed that tdTomato⁺ ECs persisted in the brain, heart, lung and liver of adults and continued to dominate the adult liver sinusoidal endothelium (Fig. 6h, j; Extended Data Figs. 9c, 10a). Accordingly, all adult organs examined contained EMP-derived ECs.

Discussion

The heterogeneous origin of blood vascular mural cells from distinct populations of mesodermal progenitors, haematopoietic and neural crest cells has been established[32]. Here we have shown that embryonic

vascular endothelium has two major origins. Thus, ECs emerge via a classical pathway of angioblast differentiation into ECs and the pathway described in this report, which entails differentiation of ECs from the EMP lineage (Extended Data Fig. 10b). Multiple previous investigations have used *Csf1r-iCre* together with recombination reporters to follow the embryonic myeloid lineage[7,10,12]. These studies predominantly used FACS with haematopoietic markers, which precluded observation of *Csf1r-iCre*-targeted ECs. By contrast, we included EC markers in FACS protocols to additionally isolate *Csf1r-iCre*-targeted ECs. In addition, immunostaining was previously used to identify *Csf1r-iCre*-targeted cells in the retina[17], liver and colon[18], but without description of EC targeting, possibly because of the close spatial proximity of ECs and perivascular macrophages[4,33]. We overcame this limitation by performing high-resolution imaging of tissues immunostained with both EC and myeloid cell markers. The contribution of EMP-derived ECs to the yolk sac, brain, heart and lung vasculature is proportionally smaller than that of ECs of classical origin, whereas EMP-derived ECs predominate in the liver, particularly the sinusoidal endothelium. Liver endothelium was previously reported to be heterogeneous in origin, with an endoderm lineage contribution of approximately 15% and the remainder of the liver EC population attributed to a venous origin[34]. Our results suggest that liver endothelium contains approximately 60% EMP-derived ECs. Preferential homing of EMPs to the liver after their entry into the embryonic circulation[13], and the dependence of liver growth on rapid vascular expansion[35], may explain the relatively large contribution of EMP-derived ECs to this organ. Ultimately, the discovery that EMPs provide a source of ECs for organ vasculature may open up new therapeutic avenues for vessel-dependent organ repair and regeneration. For example, EMPs or EMP-like EC progenitors, derived from human stem cells by modulating the expression of factors such as *Hoxa* genes, might be delivered systemically to support vascular growth in ischaemic diseases or to provide angiocrine signals that stimulate tissue stem cells.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0552-x.

Received: 16 December 2016; Accepted: 17 August 2018;
Published online 26 September 2018.

1. Potente, M., Gerhardt, H. & Carmeliet, P. Basic and therapeutic aspects of angiogenesis. *Cell* **146**, 873–887 (2011).
2. Hirschi, K. K., Ingram, D. A. & Yoder, M. C. Assessing identity, phenotype, and fate of endothelial progenitor cells. *Arterioscler. Thromb. Vasc. Biol*. **28**, 1584–1595 (2008).
3. Pollard, J. W. Trophic macrophages in development and disease. *Nat. Rev. Immunol*. **9**, 259–270 (2009).
4. Fantin, A. et al. Tissue macrophages act as cellular chaperones for vascular anastomosis downstream of VEGF-mediated endothelial tip cell induction. *Blood* **116**, 829–840 (2010).
5. Clausen, B. E., Burkhardt, C., Reith, W., Renkawitz, R. & Förster, I. Conditional gene targeting in macrophages and granulocytes using LysMcre mice. *Transgenic Res*. **8**, 265–277 (1999).
6. de Boer, J. et al. Transgenic mice with hematopoietic and lymphoid specific expression of Cre. *Eur. J. Immunol*. **33**, 314–325 (2003).
7. Hoeffel, G. et al. C-Myb$^+$ erythro-myeloid progenitor-derived fetal monocytes give rise to adult tissue-resident macrophages. *Immunity* **42**, 665–678 (2015).
8. Frame, J. M., McGrath, K. E. & Palis, J. Erythro-myeloid progenitors: "definitive" hematopoiesis in the conceptus prior to the emergence of hematopoietic stem cells. *Blood Cells Mol. Dis*. **51**, 220–225 (2013).
9. Mass, E. et al. Specification of tissue-resident macrophages during organogenesis. *Science* **353**, aaf4238 (2016).
10. Gomez Perdiguero, E. et al. Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors. *Nature* **518**, 547–551 (2015).
11. McGrath, K. E. et al. Distinct sources of hematopoietic progenitors emerge before HSCs and provide functional blood cells in the mammalian embryo. *Cell Rep*. **11**, 1892–1904 (2015).
12. Schulz, C. et al. A lineage of myeloid cells independent of Myb and hematopoietic stem cells. *Science* **336**, 86–90 (2012).
13. Ginhoux, F. & Guilliams, M. Tissue-resident macrophage ontogeny and homeostasis. *Immunity* **44**, 439–449 (2016).
14. Hoeffel, G. & Ginhoux, F. Fetal monocytes and the origins of tissue-resident macrophages. *Cell. Immunol*. **330**, 5–15 (2018).
15. Lux, C. T. et al. All primitive and definitive hematopoietic progenitor cells emerging before E10 in the mouse embryo are products of the yolk sac. *Blood* **111**, 3435–3438 (2008).
16. Fantin, A. et al. NRP1 acts cell autonomously in endothelium to promote tip cell function during sprouting angiogenesis. *Blood* **121**, 2352–2362 (2013).
17. Stefater, J. A., III et al. Regulation of angiogenesis by a non-canonical Wnt–Flt1 pathway in myeloid cells. *Nature* **474**, 511–515 (2011).
18. Deng, L. et al. A novel mouse model of inflammatory bowel disease links mammalian target of rapamycin-dependent hyperproliferation of colonic epithelium to inflammation-associated tumorigenesis. *Am. J. Pathol*. **176**, 952–967 (2010).
19. Qian, B. Z. et al. CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. *Nature* **475**, 222–225 (2011).
20. Sasmono, R. T. et al. A macrophage colony-stimulating factor receptor-green fluorescent protein transgene is expressed throughout the mononuclear phagocyte system of the mouse. *Blood* **101**, 1155–1163 (2003).
21. Burnett, S. H. et al. Conditional macrophage ablation in transgenic mice expressing a Fas-based suicide gene. *J. Leukoc. Biol*. **75**, 612–623 (2004).
22. Tam, S. J. et al. Death receptors DR6 and TROY regulate brain vascular development. *Dev. Cell* **22**, 403–417 (2012).
23. Kierdorf, K. et al. Microglia emerge from erythromyeloid precursors via Pu.1- and Irf8-dependent pathways. *Nat. Neurosci*. **16**, 273–280 (2013).
24. Goldie, L. C., Lucitti, J. L., Dickinson, M. E. & Hirschi, K. K. Cell signaling directing the formation and function of hemogenic endothelium during murine embryogenesis. *Blood* **112**, 3194–3204 (2008).
25. Wilson, C. H. et al. The kinetics of ER fusion protein activation *in vivo*. *Oncogene* **33**, 4877–4880 (2014).
26. Palis, J., Robertson, S., Kennedy, M., Wall, C. & Keller, G. Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126**, 5073–5084 (1999).
27. Alharbi, R. A., Pettengell, R., Pandha, H. S. & Morgan, R. The role of *HOX* genes in normal hematopoiesis and acute leukemia. *Leukemia* **27**, 1000–1008 (2013).
28. Toshner, M. et al. Transcript analysis reveals a specific HOX signature associated with positional identity of human endothelial cells. *PLoS ONE* **9**, e91334 (2014).
29. Rössig, L. et al. Histone deacetylase activity is essential for the expression of *HoxA9* and for endothelial commitment of progenitor cells. *J. Exp. Med*. **201**, 1825–1835 (2005).
30. Browning, A. C. et al. Comparative gene expression profiling of human umbilical vein endothelial cells and ocular vascular endothelial cells. *Br. J. Ophthalmol*. **96**, 128–132 (2012).
31. Nonaka, H., Tanaka, M., Suzuki, K. & Miyajima, A. Development of murine hepatic sinusoidal endothelial cells characterized by the expression of hyaluronan receptors. *Dev. Dyn*. **236**, 2258–2267 (2007).
32. Majesky, M. W. Developmental basis of vascular smooth muscle diversity. *Arterioscler. Thromb. Vasc. Biol*. **27**, 1248–1258 (2007).
33. Liu, C. et al. Macrophages mediate the repair of brain vascular rupture through direct physical adhesion and mechanical traction. *Immunity* **44**, 1162–1176 (2016).
34. Goldman, O. et al. Endoderm generates endothelial cells during liver development. *Stem Cell Reports* **3**, 556–565 (2014).
35. Matsumoto, K., Yoshitomi, H., Rossant, J. & Zaret, K. S. Liver organogenesis promoted by endothelial cells prior to vascular function. *Science* **294**, 559–563 (2001).

Acknowledgements We thank the Biological Resources, FACS, Imaging and Genomics facilities at UCL and E. Scarpa for technical help; D. Saur, A. Mass, D. Duboule, M. Kmita and Y. Kubota for mouse strains; and M. Golding for helpful discussions. This research was supported by grants from the Wellcome Trust (095623/Z/11/Z, 101067/Z/13/Z), Medical Research Council (MR/N011511/1) and British Heart Foundation (FS/17/23/32718).

Reviewer information *Nature* thanks L. Iruela-Arispe and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.P., A.F. and C.R. conceived and planned this study, analysed data and co-wrote the manuscript. L.D. performed genetic crosses and genotyping. A.P. and A.F. either performed experiments together or replicated each other's experiments, except for the cell cycle and *Hoxa* studies, which were carried out by A.P and A.F., respectively. J.W.P. provided mouse strains. C.R. supervised the project. All authors reviewed and edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0552-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0552-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse strains. All animal procedures were performed in accordance with the institutional Animal Welfare Ethical Review Body (AWERB) and UK Home Office guidelines. To obtain mouse embryos of defined gestational age, mice were paired in the evening and the presence of a vaginal plug the following morning was defined as E0.5. In some studies, we analysed adult mice, defined as more than eight weeks of age. Mice carrying the *Csf1r-iCre* transgene[18] were mated to mice carrying the *Cre* recombination reporters $Rosa^{Yfp}$ (ref. [36]), $Rosa^{tdTom}$ (ref. [37]) or *CAG-cat-Egfp* (ref. [38]). PU.1$^{+/-}$ mice[39] were mated to $Rosa^{Yfp}$ mice and then *Csf1r-iCre* mice to obtain *Csf1r-iCre;Rosa*^{Yfp}*;Pu.1*$^{-/-}$ embryos that lacked differentiated myeloid cells including microglia[4,39,40] as well as the myeloid cell precursors of skin pericytes[41]. $Hoxa^{fl/fl}$ mice[42] were mated to $Rosa^{tdTom}$ mice and then *Csf1r-iCre* mice to obtain *Csf1r-iCre;Rosa*^{tdTom}*;Hoxa*$^{fl/fl}$ embryos. *Csf1r-Mer-iCre-Mer* (ref. [19]) and $Kit^{CreERT2}$ (ref. [43]) as well as endothelial-specific *Cdh5-CreER*^{T2} (refs [41,44]) mice were mated to $Rosa^{tdTom}$ mice. In some experiments, mice carrying the *Csf1r-Egfp-Ngfr/Fkbp1a/Tnfrsf6* (short: *Csf1r-Egfp*) reporter of *Csf1r* expression[21] were mated to *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} mice. We also used mice with a heterozygous loss of function mutation in *Csf1* (*Csf1*$^{+/op}$)[45]. All mouse strains were maintained on a mixed background (C57Bl6/J;129/Sv), with the exception of *Csf1r-Mer-iCre-Mer*, which was maintained on a mixed FVB:C57/Bl6 background. For tamoxifen induction of CRE activity, tamoxifen (Sigma) was dissolved in peanut oil and administered via a single intraperitoneal injection into each pregnant dam. For *Csf1r-Mer-iCre-Mer* induction, we injected 1 mg tamoxifen; to achieve mosaic *Cdh5-Cre-ER*^{T2} activation, we injected 20 μg tamoxifen; for $Kit^{CreERT2}$ induction at E8.5, we injected 3 mg tamoxifen together with 1.75 mg progesterone to increase induction without inducing abortions (Sigma).

Immunolabelling. Samples were fixed in 4% formaldehyde in PBS and processed as whole-mounts or dehydrated in sucrose and embedded in optimal cutting temperature (OCT, Tissue-Tek) compound to cut 20-μm cryosections. Immunolabelling was performed as described previously for whole-mount hindbrains[46]. We used the following antibodies and dilutions: goat anti-CDH5 (1:200; AF1002, lot FQI0116101, R&D Systems), rabbit anti-CSF1R (1:500; sc-692, lot K1212, Santa Cruz), rat anti-EMCN (1:50; sc-65495, lot C2917, Santa Cruz), rabbit anti-ERG (1:200; ab92513, lot GR32027 69-1, Abcam), rat anti-F4/80 (1:500; MCA497R, lot 1605, Serotec), chicken anti-GFP (1:1,000; GFP-1020, lot 0511FP12, Aves) and rabbit anti-GFP (1:500; 598, lot 079, MBL) for YFP or EGFP labelling, rabbit anti-IBA1 (1:500; 019-19741, Wako Chemicals), rat anti-KIT (1:500; 553353, lot 30259, BD Pharmingen), rabbit anti-NG2 (1:200; AB5320, lot 2726769, Millipore), rat anti-PECAM1 (1:200; 553370, lot 5205656, BD Pharmingen), rabbit anti-pHH3 (1:400; 06-570, lot 2825969, Millipore), rabbit anti-RFP (1:1,000; PM005, lot 045, MBL), goat anti-VEGFR2 (1:200; AF644, lot COA0417021, R&D Systems). Secondary antibodies used included Alexa Fluor-conjugated goat anti-chick, -rabbit or -rat IgG (Life Technologies), or, for primary antibodies raised in goat, donkey fluorophore-conjugated FAB fragments of anti-chick, -goat, -rabbit or -rat IgG (Jackson ImmunoResearch). Note that CDH5[47], ERG[48], EMCN[49], PECAM1[50] and VEGFR2[51] were used as EC markers, whereas F4/80[52] and IBA1[53] were used as macrophage markers and NG2[54] as a pericyte marker. Biotinylated IB4 (L2140, lot 085M4032V, Sigma) followed by Alexa-conjugated streptavidin (ThermoFisher) was also used to detect brain ECs and microglia[4,16]. Nuclei were labelled with DAPI. Images were acquired with a LSM710 laser scanning confocal microscope (Zeiss) and processed using LSM image browser (Zeiss) and Photoshop CS4 (Adobe) software. Three-dimensional reconstructions including surface rendering and the generation of virtual slices for lateral views of high-resolution confocal *z*-stacks was performed using Imaris (Bitplane). *Z*-stack projections of confocal images are shown unless indicated otherwise in the figure legends.

FACS and cell culture. Tissues were mechanically and enzymatically homogenized in RPMI1640 with 2.5% fetal bovine serum (ThermoFisher), 100 μg/ml collagenase/dispase (Roche), 50 μg/ml DNase (Qiagen) and 100 μg/ml heparin (Sigma), incubated for 5 min with 0.5 mg/ml rat Fc block (Becton Dickinson) and labelled with a combination of PE/Cy7-conjugated rat anti-PECAM1 (clone 390, cat 102418, lot B212262), FITC-conjugated rat anti-CD45 (clone 30-F11, cat 103108, lot B246762) or CD41 (clone MWReg30, cat 133903, lot B201955), APC-conjugated rat anti-KIT (clone 2B8, cat 105812, lot B217855) and PerCp/Cy5.5-conjugated rat anti-CD11b (clone M1/70, cat 101227) (all BioLegend). Appropriate fluorescence gating parameters were established with unstained tissue, *Csf1r-iCre-* or *Csf1r-Egfp*-negative tissues and fluorescence-minus-one (FMO) staining. For cell cycle analysis, cell populations were incubated with 10 μg/ml Hoechst 33342 (Sigma) for 30 min at 37 °C[55] before labelling with PE/Cy7-conjugated rat anti-PECAM1 and performing FACS analysis. In all experiments, doublets were eliminated using pulse geometry gates (FSC-H versus FSC-A and SSC-H versus SSC-A), whereas dead cells were removed using SYTOX Blue (Life Technologies) or LIVE/DEAD Fixable Violet (Life Technologies). Single-cell suspensions were analysed using the BD LSRFortessa X-20 cell analyser or sorted using the BD Influx cell sorter (BD Biosciences); FlowJo software (FlowJo LLC) was used for subsequent analyses. In some experiments, a fraction of each population was cytospun onto a glass slide for Wright–Giemsa staining (Sigma) followed by imaging using an LSM510 microscope equipped with an AxioCam MRc camera (Zeiss). For cell culture experiments, cell populations were sorted into DMEM with 100 U/ml penicillin, 100 U/ml streptomycin and 20% fetal bovine serum (all ThermoFisher) before seeding the cells into a 96-well plate coated with 10 μg/ml fibronectin (ThermoFisher) to facilitate EC differentiation. Cells were then cultured in methocult (STEMCELL Technologies) to promote the formation of haematopoietic colonies, which were imaged using a TS100 microscope equipped with a DS-5M colour camera (Nikon). After removal of methocult, adherent cells were fixed with 4% formaldehyde in PBS and then labelled for VEGFR2, ERG, CD45, F4/80 and CSF1R (see above) before imaging using a Ti-E microscope (Nikon).

RNA-seq. PECAM1^{+}CD45^{-}CD11b^{-}KIT^{-} ECs were isolated from E12.5 *Csf1r-iCre;Rosa*^{tdTom} embryos and divided into tdTomato^{+} and tdTomato^{-} populations with the BD Influx cell sorter before RNA was extracted with the RNeasy Micro Kit (QIAGEN). cDNA was generated and amplified using the SMART-seq V4 ultra low input RNA kit (Clontech). 100 pg of amplified cDNA per sample was used to prepare a library with the Nextera XT kit (Illumina) and run on the NextSeq 500 sequencer (lllumina). Raw sequence data were pre-processed to trim poor quality base calls and adaptor contamination using Trimmomatic v.0.36.4[56] and aligned to the mouse mm10 genome with STAR v.2.5.2b[57]. Mapped reads were deduplicated to reduce PCR bias using Picard v2.7.1.1 software (http://broadinstitute.github.io/picard/), and the reads-per-transcript were then calculated using FeatureCount v1.4.6.p5 software[58]. Differential expression was performed using the BioConductor package DESeq2 via the SARTools wrapper v1.3.2.0[59].

RT–PCR. We extracted RNA from cells isolated with the BD Influx cell sorter (see above) with the RNeasy Micro Kit for cDNA synthesis with Superscript IV (ThermoFisher). Quantitative (q)RT–PCR was performed with SYBR Green on an HT7900 system (Applied Biosystems) using the following oligonucleotide pairs: *Actb* 5′-CACCACACCTTCTACAATGAG-3′ and 5′-GTCTCAAACATGATCTGGGTC-3′; *Cdh5* 5′-GATGCAGATGACCCCACTGT-3′ and 5′-AGGGCATCTTGTGTTCCAC-3′; *Csf1r* 5′-TGCGTCTACACAGTTCAGAG-3′ and 5′-ATGCTGTATATGTTCTTCGGT-3′; *Spi1* 5′-GCCATAGCGATCACTACTG-3′ and 5′-CAAGGTTTGATAAGGGAAGC-3′; *Hoxa11* 5′-TCTTTGCCTCTCTCCTTCCTT-3′ and 5′-TTGCAGACGCTTCTCTTTGTT-3′; *Evx1* 5′-GTGTGCTCTGGGCTCCTGT-3′ and 5′-GCCAGGGTGCCTTGAGAG-3; *Slc2a1* 5′-CCCCAGAAGGTTATTGAGGAGT and 5′-ACAAAGAGGCCGACAGAGAA; *Mrc1* 5′-ACTGGGCAATGCAAATGGAG and 5′- CCCTCAAAGTGCAATGGACA; *Oit3* 5′-CGTCTGCTTCCATGTCTACTG and 5′-GTGCTCACATTCATTTTCGTCA. For each oligonucleotide pair, a no-template control reaction was included.

Microarray analysis. Published microarray data were used to compare gene expression levels (normalized $\log_2$ OD) in E14.5 CD45^{-}PECAM1^{+} brain versus pooled lung and liver ECs (GSE35802)[22] and in HUVECs versus adult retinal ECs (GSE20986)[30] using GEO2R software (NCBI).

Statistical analysis. No randomization method was used, because tissues for analysis were allocated to experimental groups according to genotype, gestational age, organ or cell type. To ensure unbiased interpretation of results, the genotype and gestational age were disclosed only after data collection was complete, but the investigators were not blinded to sample origin (organ or cell type). All experiments involving two or more genotypes for comparison included littermate controls, and the minimum sample number was three. No statistical methods were used to predetermine sample size. The number of YFP^{+} ECs and YFP^{+} microglia in *Csf1r-iCre;Rosa*^{Yfp} hindbrains (Fig. 1a, b, f–h) was determined in three randomly chosen 0.72-mm^2 regions of each whole-mount labelled and flat-mounted hindbrain. For hindbrains in *Hoxa*-targeting experiments, the number of F4/80^{+} microglia (Extended Data Fig. 5) and tdTomato^{+} and IB4^{+} volume (Fig. 5b, c) were determined from confocal *z*-stacks of four randomly chosen 0.18-mm^2 regions on the lateral side of each hindbrain (Extended Data Fig. 5g). The *z*-stacks were surface rendered with Imaris (Bitplane) to obtain the F4/80^{+}, tdTomato^{+} and IB4^{+} volumes, and the F4/80^{+} volume was then subtracted from both the IB4^{+} and tdTomato^{+} total volumes to obtain the IB4^{+} EC and tdTomato^{+} EC volumes before calculating the ratio of tdTomato^{+} to IB4^{+} EC volume. To determine the number of vascular intersections in *Hoxa*-targeting experiments (Fig. 5b,d), the same confocal *z*-stacks were analysed with Imaris filament tracer after F4/80^{+} microglia were masked. For Figs. 1, 5, all counts obtained from one hindbrain were averaged to yield the value for that hindbrain. For all experiments, we calculated the mean value for at least three independent samples, where error bars represent the standard deviation of the mean (for details, see legends). Comparison of medians against means justified the use of a parametric test; to determine whether two datasets were significantly different, we therefore calculated *P* values with a two-tailed unpaired Student's *t*-test; $P < 0.05$ was considered significant. When more than two datasets were compared, we used the statistical tests indicated in the associated legends.

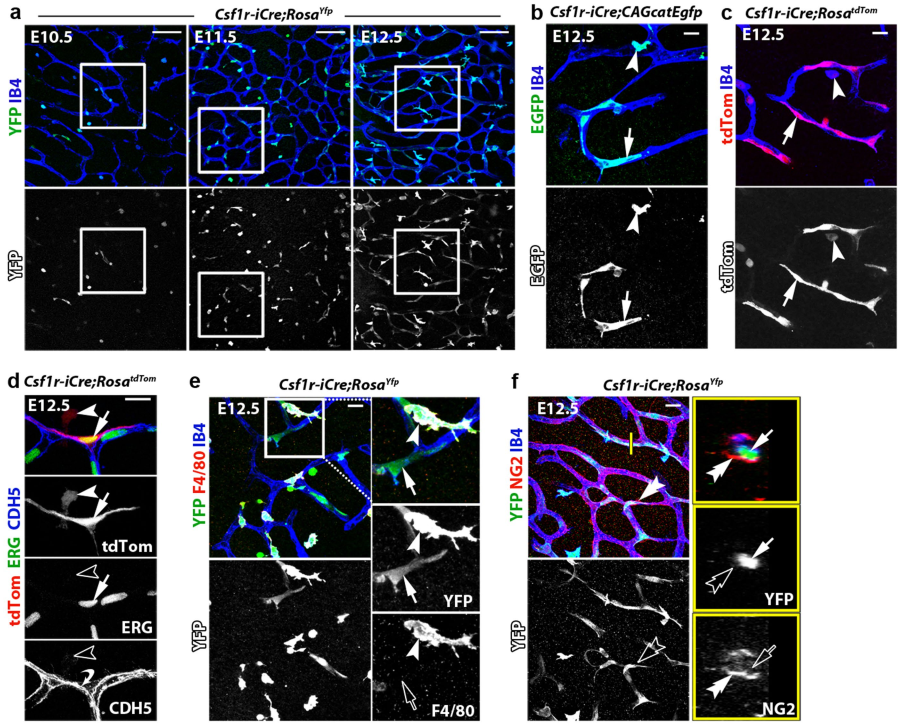
Statistical analyses were performed with Excel 12.2.6 (Microsoft Office) or Prism 7 (GraphPad Software).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

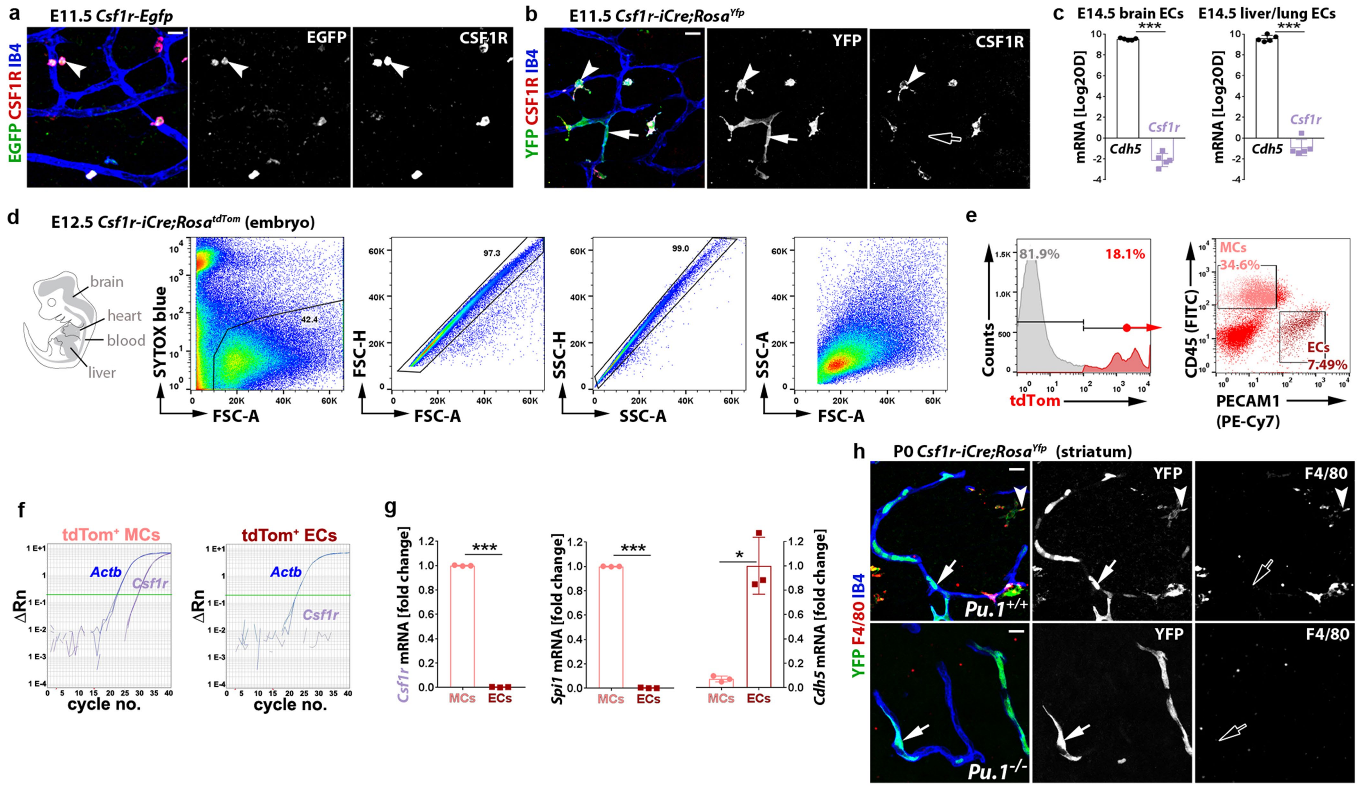
Data availability

All sequence data used in this study have been deposited in the NCBI Gene Expression Omnibus database (accession number GSE117978) and are listed in the Source Data for Fig. 6.

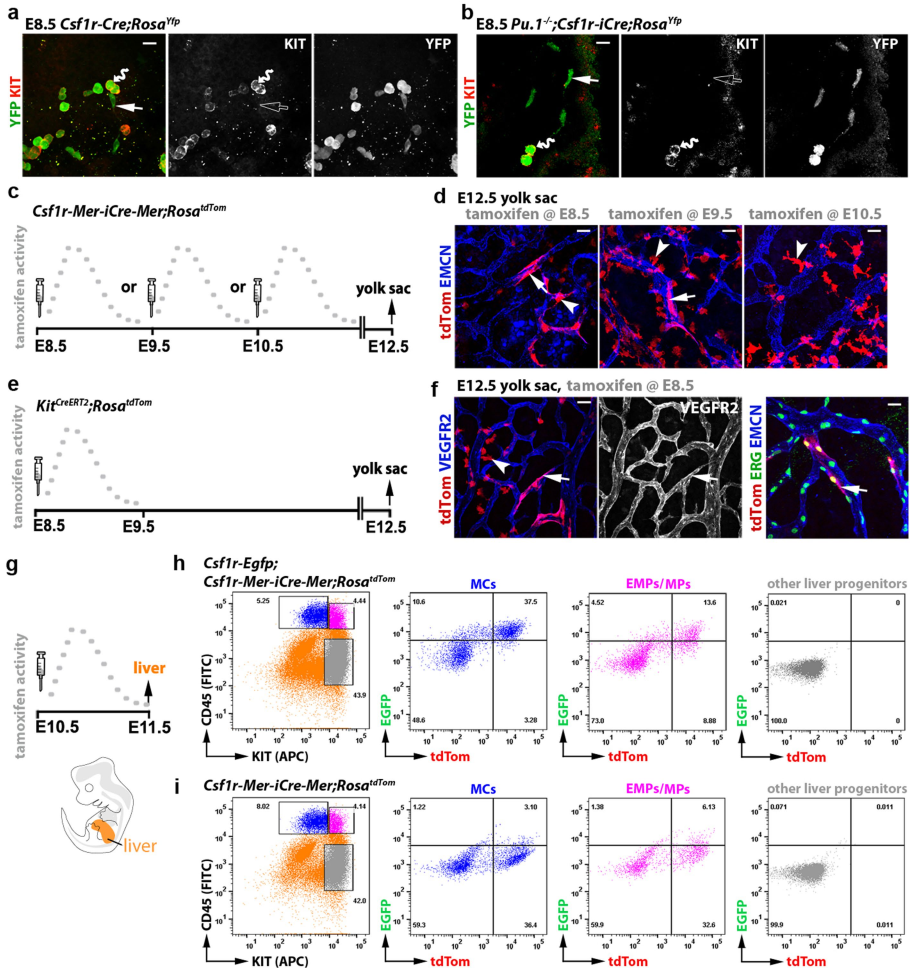
36. Srinivas, S. et al. Cre reporter strains produced by targeted insertion of *EYFP* and *ECFP* into the *ROSA26* locus. *BMC Dev. Biol*. **1**, 4 (2001).
37. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci*. **13**, 133–140 (2010).
38. Kawamoto, S. et al. A novel reporter mouse strain that expresses enhanced green fluorescent protein upon Cre-mediated recombination. *FEBS Lett*. **470**, 263–268 (2000).
39. McKercher, S. R. et al. Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *EMBO J*. **15**, 5647–5658 (1996).
40. Scott, E. W., Simon, M. C., Anastasi, J. & Singh, H. Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science* **265**, 1573–1577 (1994).
41. Yamazaki, T. et al. Tissue myeloid progenitors differentiate into pericytes through TGF-β signaling in developing skin vasculature. *Cell Rep.* **18**, 2991–3004 (2017).
42. Kmita, M. et al. Early developmental arrest of mammalian limbs lacking *HoxA*/*HoxD* gene function. *Nature* **435**, 1113–1116 (2005).
43. Klein, S. et al. Interstitial cells of Cajal integrate excitatory and inhibitory neurotransmission with intestinal slow-wave activity. *Nat. Commun*. **4**, 1630 (2013).
44. Zarkada, G., Heinolainen, K., Makinen, T., Kubota, Y. & Alitalo, K. VEGFR3 does not sustain retinal angiogenesis without VEGFR2. *Proc. Natl Acad. Sci. USA* **112**, 761–766 (2015).
45. Yoshida, H. et al. The murine mutation osteopetrosis is in the coding region of the macrophage colony stimulating factor gene. *Nature* **345**, 442–444 (1990).
46. Fantin, A., Vieira, J. M., Plein, A., Maden, C. H. & Ruhrberg, C. The embryonic mouse hindbrain as a qualitative and quantitative model for studying the molecular and cellular mechanisms of angiogenesis. *Nat. Protoc.* **8**, 418–429 (2013).
47. Gory-Fauré, S. et al. Role of vascular endothelial-cadherin in vascular morphogenesis. *Development* **126**, 2093–2102 (1999).
48. McLaughlin, F., Ludbrook, V. J., Kola, I., Campbell, C. J. & Randi, A. M. Characterisation of the tumour necrosis factor (TNF)-(alpha) response elements in the human ICAM-2 promoter. *J. Cell Sci*. **112**, 4695–4703 (1999).
49. Morgan, S. M., Samulowitz, U., Darley, L., Simmons, D. L. & Vestweber, D. Biochemical characterization and molecular cloning of a novel endothelial-specific sialomucin. *Blood* **93**, 165–175 (1999).
50. Albelda, S. M., Muller, W. A., Buck, C. A. & Newman, P. J. Molecular and cellular properties of PECAM-1 (endoCAM/CD31): a novel vascular cell-cell adhesion molecule. *J. Cell Biol*. **114**, 1059–1068 (1991).
51. Shalaby, F. et al. Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. *Nature* **376**, 62–66 (1995).
52. Austyn, J. M. & Gordon, S. F4/80, a monoclonal antibody directed specifically against the mouse macrophage. *Eur. J. Immunol*. **11**, 805–815 (1981).
53. Ohsawa, K., Imai, Y., Sasaki, Y. & Kohsaka, S. Microglia/macrophage-specific protein Iba1 binds to fimbrin and enhances its actin-bundling activity. *J. Neurochem*. **88**, 844–856 (2004).
54. Ozerdem, U., Grako, K. A., Dahlin-Huppe, K., Monosov, E. & Stallcup, W. B. NG2 proteoglycan is expressed exclusively by mural cells during vascular morphogenesis. *Dev. Dyn.* **222**, 218–227 (2001).
55. Goodell, M. A., Brose, K., Paradis, G., Conner, A. S. & Mulligan, R. C. Isolation and functional properties of murine hematopoietic stem cells that are replicating in vivo. *J. Exp. Med*. **183**, 1797–1806 (1996).
56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
57. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
58. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
59. Varet, H., Brillet-Guéguen, L., Coppée, J. Y. & Dillies, M. A. SARTools: A DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-seq data. *PLoS ONE* **11**, e0157022 (2016).



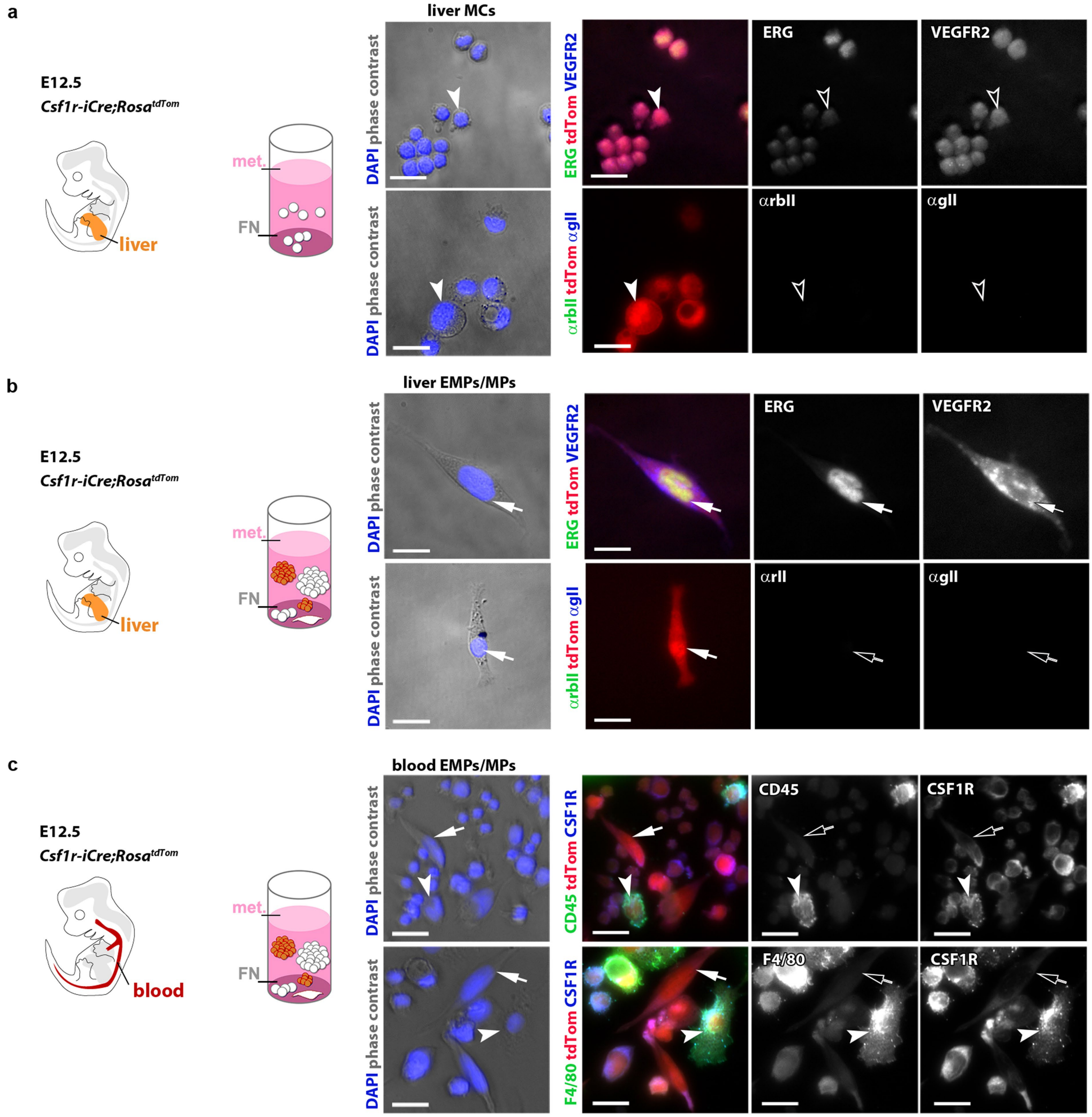
Extended Data Fig. 1 | Endothelial *Csf1r-iCre*-targeting is observed with different recombination reporters, and targeted ECs are distinguishable from macrophages and pericytes. **a**–**c**, *Csf1r-iCre;Rosa*^{Yfp} (**a**), *Csf1r-iCre;CAG-Cat-Egfp* (**b**) and *Csf1r-iCre;Rosa*^{tdTom} (**c**) hindbrains ($n = 3$ each) at the indicated stages were whole-mount labelled with IB4 and for YFP (**a**) or GFP (**b**) or are shown with tdTomato fluorescence (**c**). In **a**, the white squares indicate areas that were imaged at higher magnification for Fig. 1a. The indicated single channels are also shown individually. **d**, *Csf1r-iCre;Rosa*^{tdTom} E12.5 hindbrains ($n = 3$), whole-mount labelled for ERG and CDH5 and shown including tdTomato fluorescence to demonstrate that *Csf1r-iCre* targets ECs that form junctions with neighbouring non-targeted ECs. **e**, **f**, E12.5 *Csf1r-iCre;Rosa*^{Yfp} hindbrains, labelled for YFP and the microglia marker F4/80 (**e**) or the pericyte marker NG2 (**f**) together with IB4, show that *Csf1r-iCre*-targeted vessel-bound cells are neither microglia nor pericytes; $n = 3$ each. In **e**, the boxed area is shown in higher magnification and as single channels adjacent to the panel. In **f**, a single optical y/z cross section at the position indicated with the yellow line is displayed at higher magnification with single channels. Arrowheads, microglia; arrows, ECs; double arrowheads, pericytes; curved arrow, junctional CDH5 staining; solid and clear symbols indicate the presence or absence of marker expression, respectively. Scale bars: 100 μm (**a**), 20 μm (**b**, **c**, **e**, **f**), 50 μm (**d**).



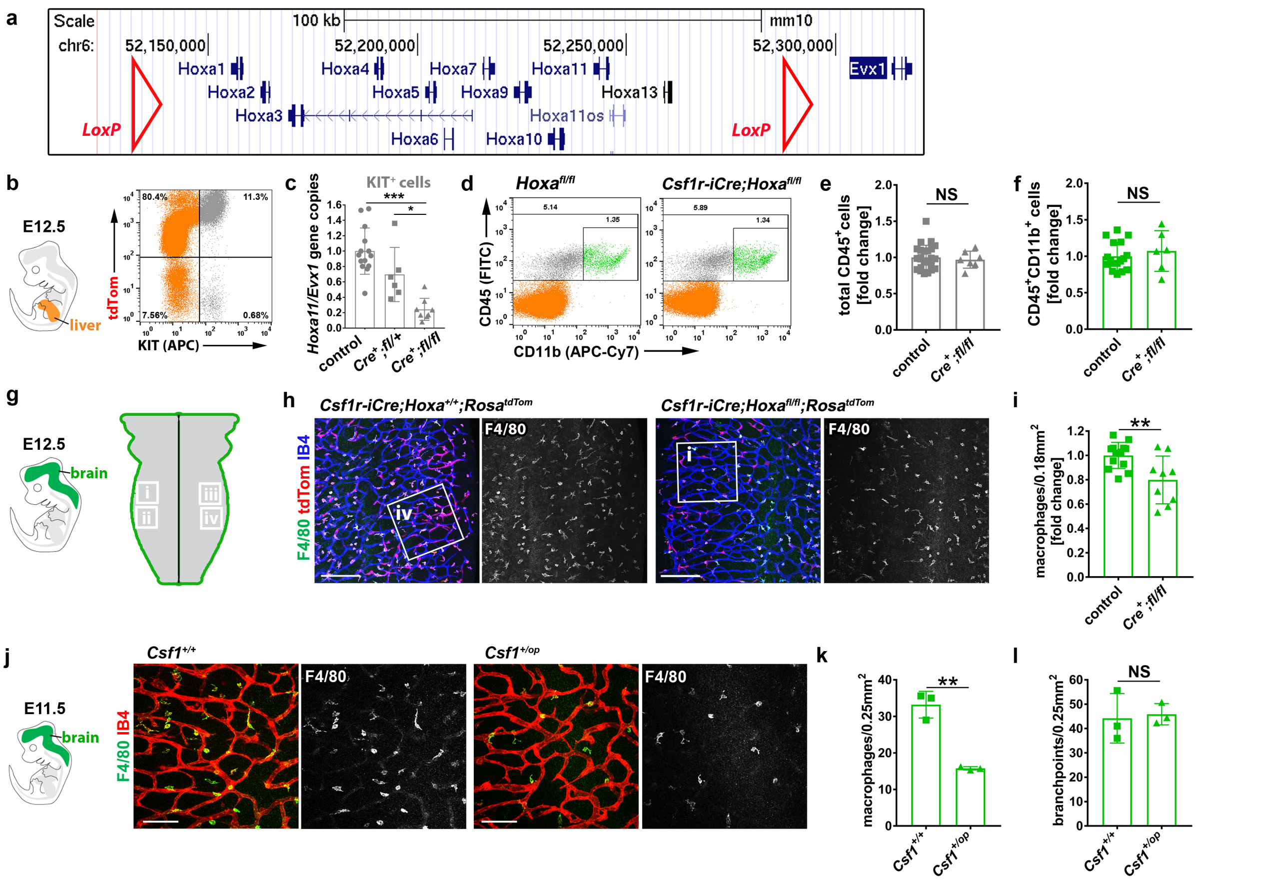
Extended Data Fig. 2 | Endothelial *Csf1r-iCre*-targeting is not caused by endothelial *Csf1r* expression and occurs independently of myeloid differentiation. a, **b**, *Csf1r-Egfp* (**a**) and *Csf1r-iCre;Rosa*^{Yfp} (**b**) E11.5 hindbrains ($n = 3$ each), whole-mount labelled for CSF1R and EGFP or YFP together with IB4, show lack of *Csf1r* promoter activity and CSF1R protein in ECs. **c**, Relative *Cdh5* and *Csf1r* expression levels in our analysis of published E14.5 brain or pooled lung/liver EC microarrays[22]; $n = 5$ each; $***P < 0.0001$ (two-tailed unpaired *t*-test). **d**–**g**, FACS separation of tdTomato^{+} cells from *Csf1r-iCre;Rosa*^{tdTom} embryos ($n = 3$) for gene expression analysis, including representative gating strategy to exclude dead cells and doublets in this and subsequent experiments (**d**) and sorting into PECAM1^{+}CD45^{-} ECs versus CD45^{+}PECAM1^{-} MCs (**e**). **f**, Representative RT–qPCR gene amplification graphs for *Csf1r* and *Actb* from tdTomato^{+} MCs and ECs; ΔRn, normalized reporter value for SYBR Green minus baseline instrument signals. **g**, Graphic representation of the fold change in RT–qPCR amplification of the indicated genes relative to *Actb* for both cell populations; each data point represents one embryo; $*P = 0.0242$, $***P < 0.0001$ (two-tailed unpaired *t*-test). **h**, *Csf1r-iCre;Rosa*^{Yfp} P0 striatum on a *Pu.1*$^{+/+}$ versus *Pu.1*$^{-/-}$ background ($n = 3$ brains each), cryosectioned and labelled for YFP and F4/80 together with IB4 to show that *Csf1r-iCre*-targeted ECs are PU.1-independent and persist postnatally. Arrowheads, microglia; arrows, YFP^{+} ECs; clear arrows, YFP^{+} ECs that are CSF1R^{-} and F4/80^{-}. Scale bars, 20 μm.



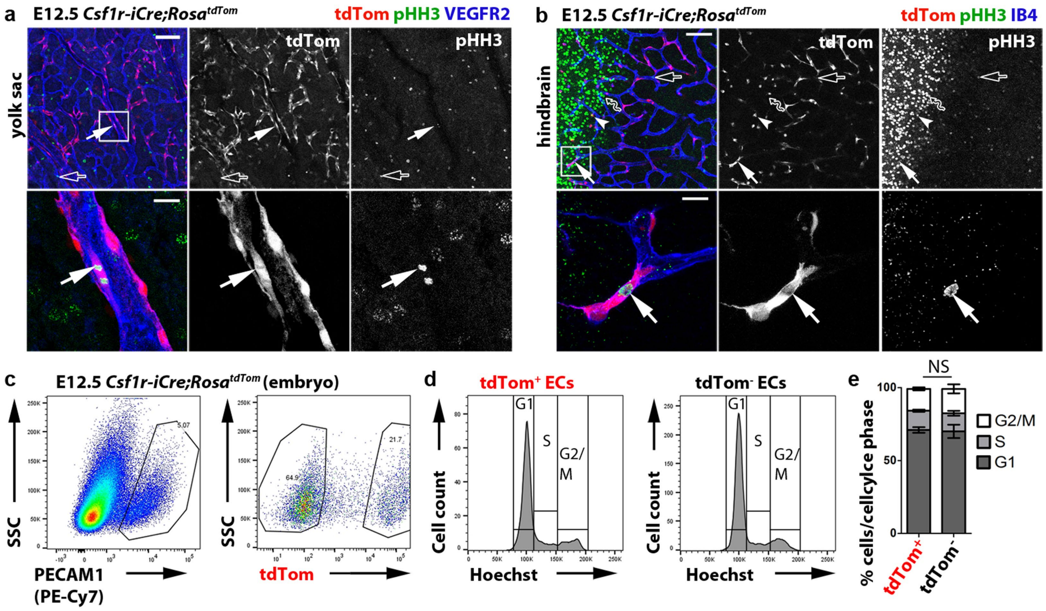
Extended Data Fig. 3 | Lineage tracing of yolk sac and liver EMPs. **a**, **b**, E8.5 wild-type (**a**) and *Pu.1*$^{-/-}$ (**b**) yolk sacs on a *Csf1r-iCre;Rosa*^{Yfp} background ($n = 3$ yolk sacs each), whole-mount labelled for YFP and KIT, show *Csf1r-iCre*-targeted KIT^{+} round cells corresponding to EMPs and MPs as well as *Csf1r-iCre*-targeted KIT^{-} flat cells corresponding to ECs. Scale bars, 20 μm. **c**–**f**, Pregnant *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} (**c**, **d**) and *Kit*^{CreERT2}*;Rosa*^{tdTom} (**e**, **f**) dams were injected with a single tamoxifen dose on the indicated days; E12.5 yolk sacs were whole-mount labelled for the indicated markers to identify *Csf1r-iCre*-targeted ECs and macrophages ($n = 3$ yolk sacs for each genotype). Wavy arrows, EMPs; straight arrows, *Csf1r-iCre*-lineage-traced ECs; arrowheads, macrophages; solid and clear symbols indicate the presence or absence, respectively, of the indicated markers. Scale bars, 20 μm. **g**–**i**, Pregnant dams were injected with a single tamoxifen dose on E10.5 (**g**) before using the indicated markers for FACS analysis of E11.5 *Csf1r-Egfp;Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} (**h**) or *Csf1r-Mer-iCre-Mer;Rosa*^{tdTom} control (**i**) livers ($n = 4$ each); the CD45^{high}KIT^{-} differentiated MC (blue), CD45^{low}KIT^{+} EMP/MP (pink) and CD45^{-}KIT^{+} (grey) populations were gated further for *Csf1r-Egfp* and tdTomato. CD45^{-}KIT^{+} cells were neither MCs nor EMPs, because they lacked CD45, tdTomato and EGFP.



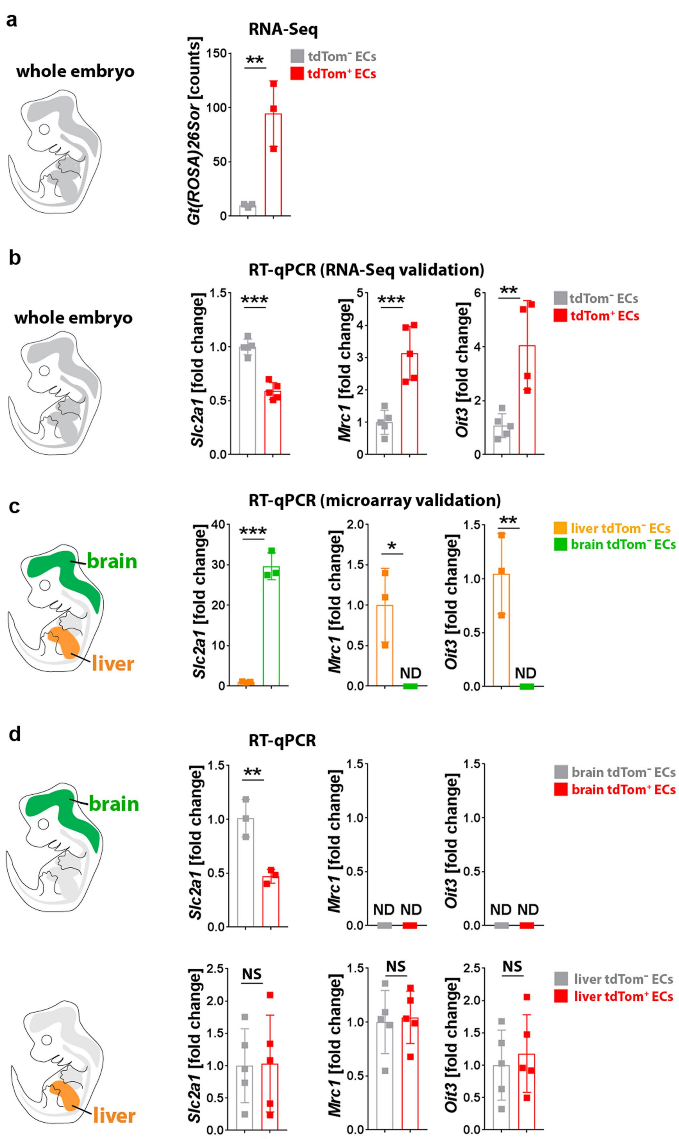
Extended Data Fig. 4 | Immunostaining controls for cultured *Csf1r-iCre*-targeted cells. The indicated cell populations were FACS-isolated from E12.5 *Csf1r-iCre;Rosa*^{tdTom} liver or blood with the indicated markers and cultured for three days in methocult (met.) on fibronectin (FN); $n = 1$ experiment. **a**, **b**, Adherent cells from tdTomato^{+} liver MC (**a**) and EMP/MP (**b**) cultures were stained for ERG and VEGFR2 (top) or with secondary antibodies only (bottom). **c**, Adherent cells from tdTomato^{+} blood EMP and MP cultures were immunostained for CSF1R together with the myeloid markers CD45 (top) or F4/80 (bottom). In the first panel in each row, the phase contrast and DAPI images were merged. In panels 2–4 in each row, immunolabelled cells were visualized together with tdTomato fluorescence, with single channels for the indicated markers shown separately in greyscale. Arrows, tdTomato^{+} ECs; arrowheads, tdTomato^{+} MCs; solid and clear symbols indicate the presence or absence, respectively, of the indicated markers. Scale bars, 20 μm.



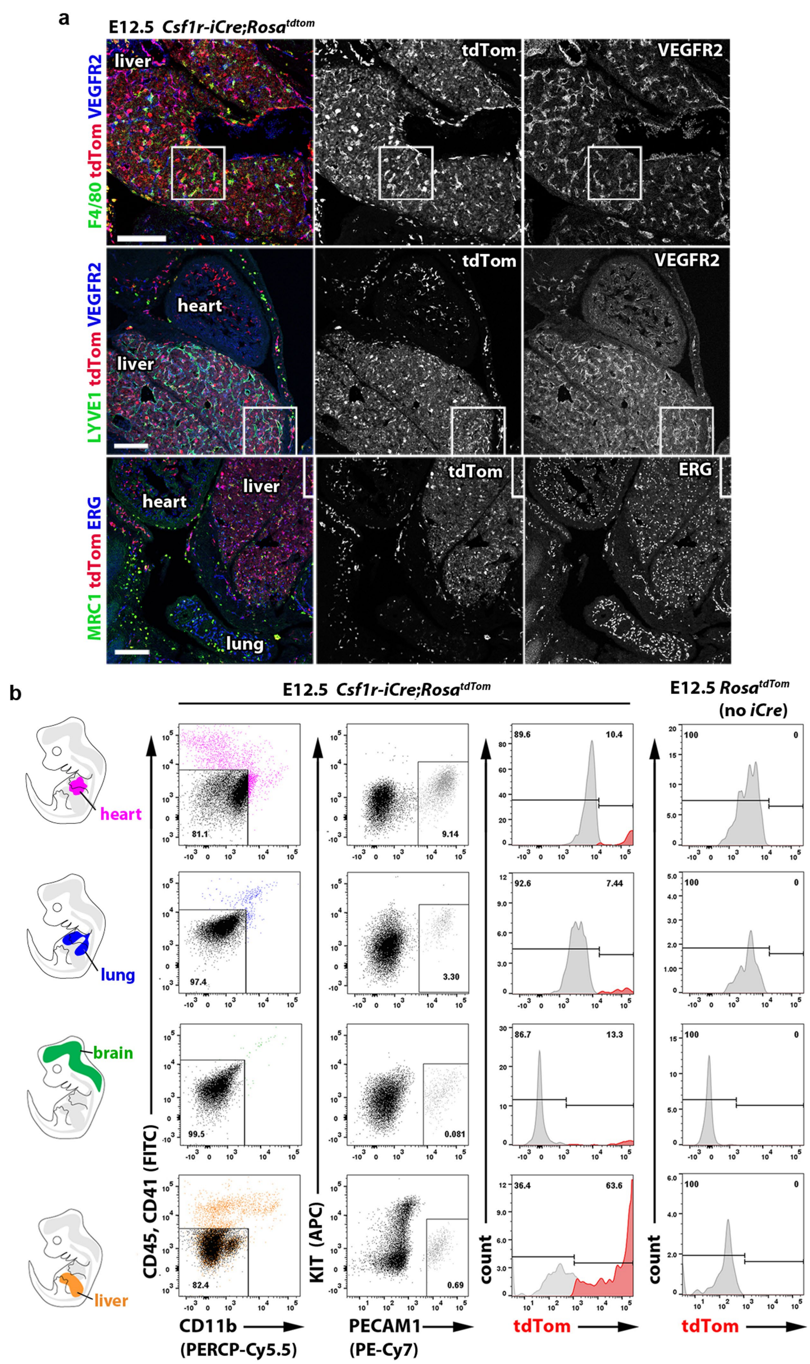
Extended Data Fig. 5 | *Hoxa* gene targeting with *Csf1r-iCre*. **a**, Schematic representation of the *Hoxa* gene cluster and adjacent *Evx1* gene using the UCSC Genome Browser with the mouse December 2011 (GRCm38/mm10) Assembly, including position of the *LoxP* sites used for gene targeting. **b**, **c**, Validation of *Hoxa* targeting. **b**, FACS strategy to isolate KIT$^+$ cells from E12.5 control (pooled *Csf1r-iCre*$^-$ *or Csf1r-iCre*$^+$*;Hoxa*$^{+/+}$; $n = 14$), *Hoxa*$^{+/fl}$*;Csf1r-iCre* ($n = 6$) and *Hoxa*$^{fl/fl}$*;Csf1r-iCre* ($n = 8$) livers. **c**, qPCR analysis of *Hoxa11* gene copy number relative to *Evx1*; mean ± s.d.; each symbol represents the value for one liver; $^*P = 0.0156$, $^{***}P < 0.001$ (one-way ANOVA, Tukey's multiple comparisons test). **d**–**f**, Representative FACS analysis (**d**) and quantification (**e**, **f**) of liver cell populations at E12.5 shows a similar number of total CD45$^+$ and CD45$^+$CD11b$^+$ differentiated myeloid cells in *Hoxa*$^{fl/fl}$*;Csf1r-iCre* mutants ($n = 7$ for CD45$^+$; $n = 6$ for CD45$^+$CD11b$^+$) versus pooled *Csf1r-iCre*$^-$ and *Csf1r iCre*$^+$*;Hoxa*$^{+/+}$controls ($n = 25$ for CD45$^+$, $n = 17$ for CD45$^+$CD11b$^+$); mean ± s.d. fold change in mutants compared to controls; each data point represents one liver; NS, not significant, $P = 0.6519$ (**e**) and $P = 0.496$ (**f**) (two-tailed unpaired *t*-test). **g**–**i**, E12.5 hindbrains of the indicated genotypes were immunolabelled to determine vascular complexity and quantify microglia. **g**, Schematic representation of a whole-mount embryonic hindbrain (left) and location of the hindbrain areas i–iv used for quantification (right); values for the four areas in each hindbrain were averaged to obtain the value for that hindbrain; EC quantifications are shown in Fig. 5c. **h**, Hindbrains were whole-mount labelled with IB4 and for RFP to visualize tdTomato and for F4/80 to visualize microglia; white boxes indicate areas shown in higher magnification in Fig. 5. **i**, Quantification of microglia in *Hoxa*$^{fl/fl}$*;Csf1r-iCre* mutants ($n = 9$) versus controls ($n = 10$, pooled *Csf1r-iCre*$^+$*;Hoxa*$^{+/+}$ and *Csf1r-iCre*$^-$ of any *Hoxa* genotype); mean ± s.d. fold change in mutant compared to control hindbrain; each data point represents one hindbrain; $^{**}P = 0.0055$ (two-tailed unpaired *t*-test). **j**–**l**, E11.5 *Csf1*$^{+/+}$ and *Csf1*$^{+/op}$ littermate hindbrains, whole-mount labelled for F4/80 together with IB4 (**j**) before quantification of microglia number (**k**) and vascular branchpoints as a measure of vascular complexity (**l**). Mean ± s.d.; each data point represents one hindbrain, $n = 3$ each; NS, not significant, $P = 0.808$, $^{**}P = 0.0012$ (two-tailed unpaired *t*-test). Scale bars: 200 μm (**h**), 100 μm (**j**).



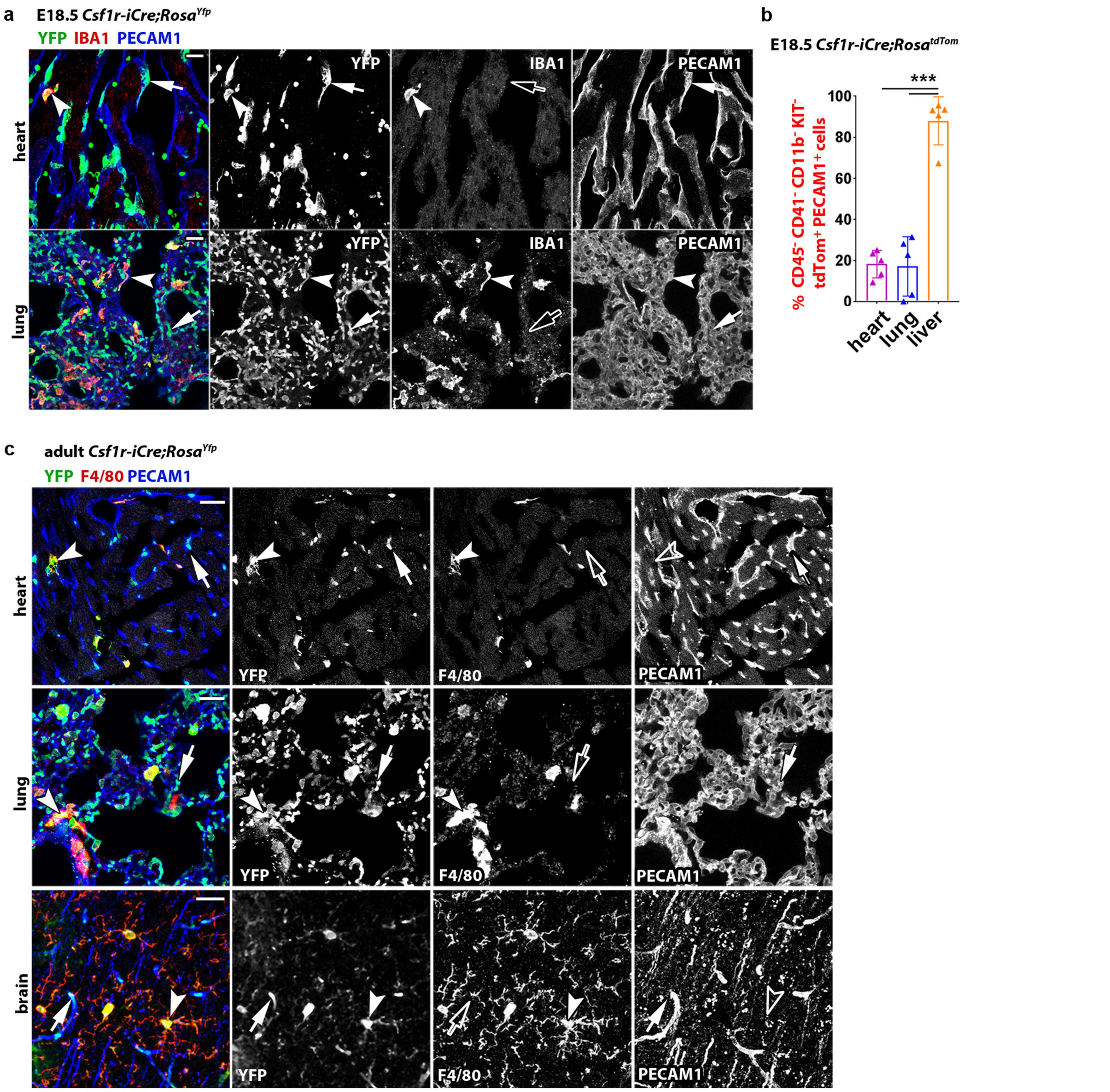
Extended Data Fig. 6 | *Csf1r-iCre*-targeted ECs proliferate in vivo. **a**, **b**, E12.5 *Csf1r-iCre;Rosa*^{tdTom} yolk sac (**a**) or hindbrain (**b**), wholemount stained for the proliferation marker pHH3 and VEGFR2 or for pHH3 together with IB4, respectively, and shown together with tdTomato fluorescence ($n = 3$ each). Areas indicated with white squares were imaged at higher magnification and are shown below the corresponding panels, with tdTomato and pHH3 channels also shown separately in greyscale. Arrows, proliferating tdTomato$^+$pHH3$^+$ ECs; solid and clear symbols indicate the presence or absence, respectively, of tdTomato fluorescence; wavy arrow, a tdTomato$^-$pHH3$^+$ neural progenitor. Scale bars: 100 μm (top), 20 μm (bottom). **c**–**e**, Cell cycle distribution of tdTomato$^+$ and tdTomato$^-$ ECs. **c**, FACS strategy to isolate tdTomato$^+$ and tdTomato$^-$PECAM1$^+$ ECs from E12.5 *Csf1r-iCre;Rosa*^{tdTom} embryos ($n = 3$ embryos). **d**, Cell cycle distribution based on Hoechst 33342 fluorescence as a measure of DNA content; low and high staining intensity is observed in cells with a DNA ploidy of $2n$ (G0/G1 phase) or $4n$ (G2/M phase), respectively; intermediate staining intensity corresponds to S phase. **e**, Mean ± s.d. proportion of tdTomato$^+$ and tdTomato$^-$ ECs in G1, S and G2/M based on the area of the corresponding peaks in **d**; NS, not significant, $P > 0.9999$ (two-way ANOVA, Bonferroni's multiple comparisons test).



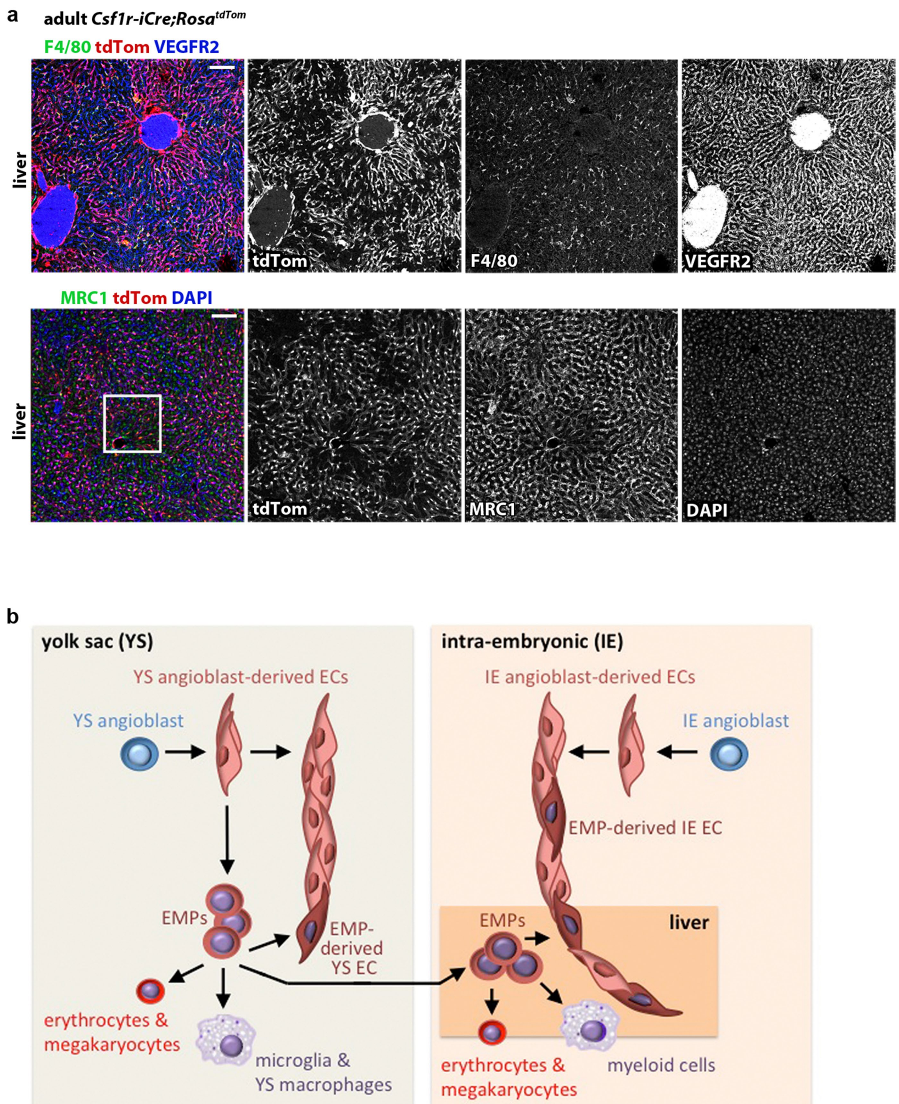
Extended Data Fig. 7 | Validation of gene expression data from RNA-Seq and microarray studies. ECs were FACS-isolated from E12.5 *Csf1r-iCre;Rosa*^{tdTom} embryos ($n = 3$) as in Fig. 6a to validate the RNA-seq and microarray data shown in Fig. 6d–f. *Slc2a1* was analysed as a representative brain EC-enriched transcript/differentiation marker, and *Mrc1* and *Oit3* as representative liver EC-enriched transcripts. **a**, Relative transcript levels of the *Gt(ROSA)26Sor* (*tdTomato*) transcript by RNA-seq of the E12.5 tdTomato$^+$ and tdTomato$^-$ EC populations (analysis presented in Fig. 6a–f); mean ± s.d. of normalized counts, $n = 3$ each; $^{**}P = 0.0085$ (two-sided unpaired *t*-test). **b**, RT–qPCR analysis for the indicated genes in tdTomato$^+$ versus tdTomato$^-$ ECs isolated from whole E12.5 embryos ($n = 5$) to validate genes identified by RNA-seq in Fig. 6e, f as differentially expressed. Mean ± s.d. of fold change; $^{***}P < 0.0001$ (*Slc2a1*), $^{***}P = 0.0008$ (*Mrc1*) $^{**}P = 0.0056$ (*Oit3*) (two-sided unpaired *t*-test). **c**, RT–qPCR analysis for the indicated genes in tdTomato$^-$ ECs isolated from the E12.5 brain versus liver ($n = 3$ for each organ) to validate organ-specific transcript enrichment identified via microarray analysis shown in Fig. 6f. Mean ± s.d. of fold change; $^{*}P = 0.019$, $^{**}P = 0.0082$, $^{***}P < 0.0001$ (two-sided unpaired *t*-test); ND, not detectable. **d**, RT–qPCR analysis for the indicated genes to directly compare the expression levels of brain and liver EC differentiation markers in tdTomato$^+$ versus tdTomato$^-$ ECs isolated from brain ($n = 3$) or liver ($n = 5$). Mean ± s.d. of fold change; NS, not significant, $P = 0.9398$ (liver *Slc2a1*), $P = 0.8045$ (liver *Mrc1*), $P = 0.6327$ (liver *Oit3*), $^{**}P = 0.0073$ (brain *Slc2a1*) (two-sided unpaired *t*-test); ND, not detectable.



Extended Data Fig. 8 | *Csf1r-iCre*-targeted ECs contribute to embryonic vasculature in multiple organs. a, 20-μm cryosections of the indicated E12.5 *Csf1r-iCre;Rosa*^{tdTom} organs ($n = 3$ each) were immunolabelled for the indicated EC markers together with antibodies for RFP to identify tdTomato protein (top and bottom) or are shown with tdTomato fluorescence (middle); single channels are shown in greyscale. The white boxes indicate the positions of areas shown in higher magnification in Fig. 6g; some areas selected for higher magnification are not contained entirely within the field of view, and accordingly the boxes are shown incomplete. Scale bars, 200 μm. **b**, Gating strategy for FACS analysis of tdTomato^{+} and tdTomato^{-} ECs from E12.5 *Csf1r-iCre;Rosa*^{tdTom} brain, lung, heart and liver versus control organs lacking *iCre*, using antibodies for CD11b, CD41, CD45, KIT and PECAM1; associated EC quantifications are shown in Fig. 6i. An analogous strategy was used for the quantifications shown in Fig. 6j and in Extended Data Fig. 9b.



Extended Data Fig. 9 | *Csf1r-iCre*-targeted ECs contribute to organ vasculature in late-stage embryos and adults. a, 20-μm cryosections of the indicated organs from E18.5 *Csf1r-iCre;Rosa*^{Yfp} mice ($n = 2$ each) were immunolabelled for YFP, PECAM1 and IBA1; single channels are shown in greyscale. Arrowheads, YFP^{+}IBA1^{+} macrophages; solid and empty arrows, ECs that are YFP^{+} and lack IBA1 expression, respectively. Scale bars, 20 μm. **b**, FACS analysis of dissociated cells from the indicated organs of E18.5 *Csf1r-iCre;Rosa*^{tdTom} embryos after staining with antibodies for CD11b, CD41, CD45, KIT and PECAM1, using the gating strategy shown in Extended Data Fig. 8b; mean ± s.d., $n = 5$ each; ***$P < 0.0001$ (one-way ANOVA, Tukey's multiple comparisons test). **c**, 20-μm cryosections of the indicated organs from 6-month-old adult *Csf1r-iCre;Rosa*^{Yfp} mice ($n = 3$ organs each) were immunolabelled for YFP, PECAM1 and F4/80; single channels are shown in greyscale. Arrowheads and arrows as in **a**. Scale bars, 20 μm.



Extended Data Fig. 10 | *Csf1r-iCre*-targeted ECs contribute to adult organ vasculature. a, 20-μm cryosections of 3-month-old adult *Csf1r-iCre;Rosa*^{tdTom} livers ($n = 3$) were immunolabelled for RFP, VEGFR2 and F4/80 or MRC1 and then counterstained with DAPI; single channels are shown in greyscale. The white box indicates an area shown in higher magnification in Fig. 6h. Scale bars, 100 μm. **b**, Working model for the role of EMPs in generating extra-embryonic yolk sac and intra-embryonic organ ECs alongside their known role in generating myeloid and erythrocyte/megakaryocyte lineage cells. It is not yet known whether EMP-derived and non-EMP-derived ECs have different functions to regulate normal organ physiology or pathological vascular responses in the adult.

https://doi.org/10.1038/s41586-018-0564-6

Nearly all the sky is covered by Lyman-α emission around high-redshift galaxies

L. Wisotzki[1]*, R. Bacon[2], J. Brinchmann[3,4], S. Cantalupo[5], P. Richter[6], J. Schaye[3], K. B. Schmidt[1], T. Urrutia[1], P. M. Weilbacher[1], M. Akhlaghi[3], N. Bouché[7], T. Contini[7], B. Guiderdoni[2], E. C. Herenz[8], H. Inami[2], J. Kerutt[1], F. Leclercq[2], R. A. Marino[5], M. Maseda[3], A. Monreal-Ibero[9,10], T. Nanayakkara[3], J. Richard[2], R. Saust[1], M. Steinmetz[1] & M. Wendt[1,6]

Galaxies are surrounded by large reservoirs of gas, mostly hydrogen, that are fed by inflows from the intergalactic medium and by outflows from galactic winds. Absorption-line measurements along the lines of sight to bright and rare background quasars indicate that this circumgalactic medium extends far beyond the starlight seen in galaxies, but very little is known about its spatial distribution. The Lyman-α transition of atomic hydrogen at a wavelength of 121.6 nanometres is an important tracer of warm (about 10^4 kelvin) gas in and around galaxies, especially at cosmological redshifts greater than about 1.6 at which the spectral line becomes observable from the ground. Tracing cosmic hydrogen through its Lyman-α emission has been a long-standing goal of observational astrophysics[1–3], but the extremely low surface brightness of the spatially extended emission is a formidable obstacle. A new window into circumgalactic environments was recently opened by the discovery of ubiquitous extended Lyman-α emission from hydrogen around high-redshift galaxies[4,5]. Such measurements were previously limited to especially favourable systems[6–8] or to the use of massive statistical averaging[9,10] because of the faintness of this emission. Here we report observations of low-surface-brightness Lyman-α emission surrounding faint galaxies at redshifts between 3 and 6. We find that the projected sky coverage approaches 100 per cent. The corresponding rate of incidence (the mean number of Lyman-α emitters penetrated by any arbitrary line of sight) is well above unity and similar to the incidence rate of high-column-density absorbers frequently detected in the spectra of distant quasars[11–14]. This similarity suggests that most circumgalactic atomic hydrogen at these redshifts has now been detected in emission.

We used the Multi-Unit Spectroscopic Explorer (MUSE), developed by our team and installed in 2014 at the ESO Very Large Telescope[15–17], to perform very long exposures in two fields that were previously mapped to extreme depths with the Hubble Space Telescope (HST): the Hubble Deep Field South[18] (HDFS) and the Hubble Ultra Deep Field[19] (HUDF). In our MUSE data we detected 270 Lyman-α (Lyα)-emitting galaxies at $3 < z < 6$ (Methods), many of which are barely visible or even undetected with the HST. Extended Lyα haloes around these galaxies can be traced to distances of a few arcseconds from the source centres[4,5].

In a first approach to estimating the total sky coverage of extended Lyα emission, we constructed redshift-integrated Lyα maps in the two observed fields as follows (Methods): we extracted pseudo-narrowband Lyα subimages around each selected object from the MUSE data. Coadding all subimages over the full redshift range followed by some spatial filtering yielded a Lyα image for each field. Figure 1 shows this image for the HUDF. Counting the number of pixels above a given Lyα surface brightness $s_{\mathrm{Ly}\alpha}$ yields the fractional sky coverage $f_{\mathrm{Ly}\alpha}$. For a threshold of $s_{\mathrm{Ly}\alpha} > 10^{-19}$ erg s^{-1} cm^{-2} arcsec^{-2}—the typical 1σ limiting surface brightness in the narrowband images inside an aperture of 1″—we find a sky coverage of 46% in the HUDF and 45% in the HDFS (Methods). Although this result already suggests that the sky coverage might further increase for even lower thresholds, the approach is hampered by noise and the need to apply spatial filtering.

To lower the surface brightness limit beyond the sensitivity of individual lines of sight, we employed a combination of azimuthal averaging and image stacking (Methods). We first computed radial Lyα surface brightness profiles for each object by averaging over pre-defined concentric annuli. Motivated by the fact that our Lyα halo profiles do not, on average, depend strongly on Lyα luminosity[5], we median-combined the individual images in three redshift bins (Methods) and approximated the radial profiles of the median images by smooth fitting functions. Figure 2 illustrates this process. In the outermost annuli,

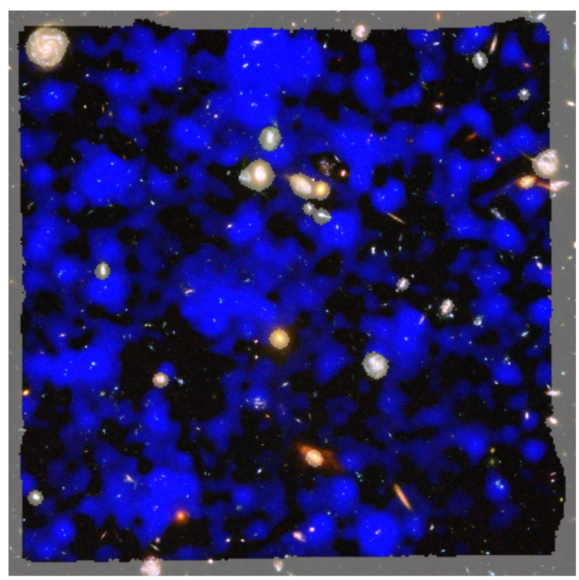


Fig. 1 | Distribution of the observed Lyα emission in the HUDF. The underlying image is a colour composite obtained by the HST[19] restricted to the 1′ × 1′ section observed with MUSE. The extended Lyα emission detected by MUSE is superimposed in blue, summed over the redshift range $3 < z < 6$ and spatially filtered to suppress the noise. The grey semi-transparent areas outline the MUSE field of view and also mask the brightest foreground galaxies. The dynamic range of the Lyα overlay was adjusted such that the faintest visible structures have a surface brightness of 10^{-19} erg s^{-1} cm^{-2} arcsec^{-2}. Credit: NASA, ESA, and S. Beckwith (STScI) and the HUDF Team (CC BY 4.0; https://creativecommons.org/licenses/by/4.0/); crop, alignment, and application of blur filter by L.W.

[1]Leibniz-Institut für Astrophysik Potsdam (AIP), Potsdam, Germany. [2]Université Lyon, Université Lyon 1, ENS de Lyon, CNRS, Centre de Recherche Astrophysique de Lyon, Saint-Genis-Laval, France. [3]Leiden Observatory, Leiden University, Leiden, The Netherlands. [4]Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Porto, Portugal. [5]Department of Physics, ETH Zürich, Zürich, Switzerland. [6]Institut für Physik und Astronomie, Universität Potsdam, Potsdam, Germany. [7]Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, Toulouse, France. [8]Department of Astronomy, Stockholm University, AlbaNova University Centre, Stockholm, Sweden. [9]Instituto de Astrofísica de Canarias (IAC), La Laguna, Tenerife, Spain. [10]Universidad de La Laguna, Departamento Astrofísica, La Laguna, Tenerife, Spain. *e-mail: lwisotzki@aip.de

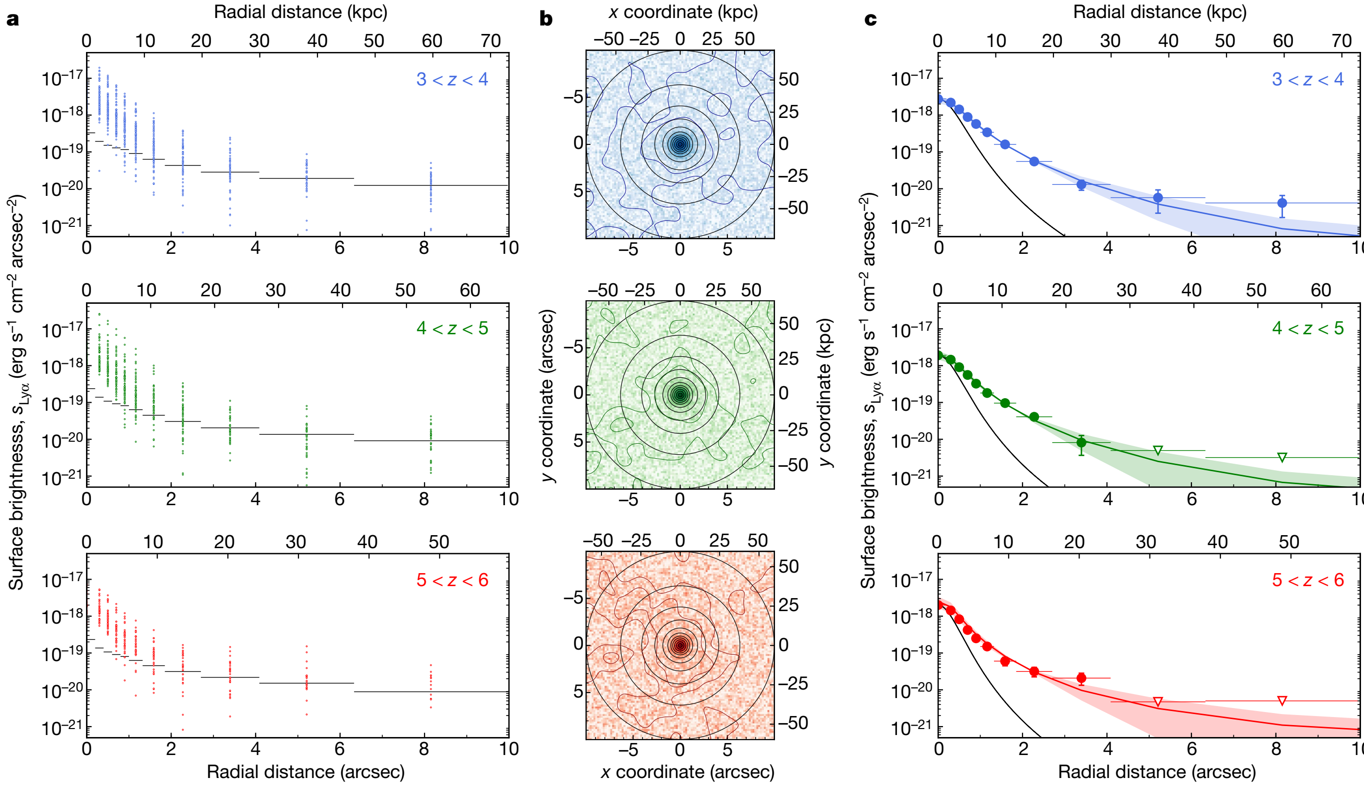


Fig. 2 | Stacking and construction of representative radial Lyα profiles. **a**, Lyα surface brightness measurements of the individual galaxies in the HUDF and HDFS, azimuthally averaged in concentric annuli (dots). The surface brightness levels indicated by the horizontal error bars correspond to the typical 1σ uncertainties, and their horizontal extents give the widths of the annuli. The three panels (top, middle and bottom) represent three disjoint redshift ranges as given by the labels at top right of the panels. Radial coordinates are given in angular (bottom axes) as well as physical (top axes) units, the latter evaluated at the centre of each redshift range assuming a cosmological model with $h = 0.7$, $\Omega_{\mathrm{m}} = 0.3$ and $\Omega_{\Lambda} = 0.7$. **b**, Median-stacked Lyα images for these redshift bins (indicated by colours to match those in **a**). The contours trace surface brightnesses of $(0.5, 2) \times 10^{-20}$ erg s^{-1} cm^{-2} arcsec^{-2} after subtracting a model image, smoothing the residual with a Gaussian of 2″ full-width at half-maximum (FWHM) and adding back the model. The overplotted black circles show the boundaries of the annuli used to extract the radial profiles. **c**, Azimuthally averaged radial profiles of the median-stacked images (data points), with inverted triangles indicating upper limits. The vertical bars on the data points quantify the estimated 1σ errors, while the horizontal bars again indicate the widths of the annuli. The black line in each panel traces the radial shape of a scaled point source[4,5], demonstrating that the median Lyα emission is well resolved for radii greater than about 1 arcsec. The solid, coloured curves show the profiles extracted from two-dimensional surface brightness model fits to the median images, with the shaded regions indicating the estimated 1σ uncertainties.

the median-stacked profiles reach limiting surface brightness levels of $s_{\mathrm{Ly}\alpha,\mathrm{lim}} \approx (5, 4, 4) \times 10^{-21}$ erg s^{-1} cm^{-2} arcsec^{-2} (1σ, at $z \approx 3.5, 4.5, 5.5$), an order of magnitude more sensitive than for the single line of sight measurements considered above.

From the scaled median-stacked profiles we created synthetic Lyα maps for the three redshift bins and for the full redshift range. These maps, shown in Fig. 3a, represent idealized, noise-free and approximately seeing-corrected models of the Lyα distribution in the sky. The resulting cumulative fractional Lyα sky coverage $f_{\mathrm{Ly}\alpha}$ is presented in Fig. 3b as a function of the surface brightness threshold. At $s_{\mathrm{Ly}\alpha} \approx 10^{-20}$ erg s^{-1} cm^{-2} arcsec^{-2}, $f_{\mathrm{Ly}\alpha}$ is already well above 80% and still increasing. At these extremely faint levels, the contributions of $f_{\mathrm{Ly}\alpha}$ from the different redshift ranges formally add up to more than 100%, a clear sign that the Lyα emission regions substantially overlap in projection.

The sky coverage is an intuitively appealing number but of limited use as it saturates at 100%. A closely related but more physically useful quantity is the incidence rate dn/dz, the average number of Lyα-emitting regions per unit redshift passed by a typical line of sight, at a given surface brightness level. This quantity can be directly compared to dn/dz (obtained from absorption line statistics) for different absorber column densities. We also corrected for cosmological surface brightness dimming by moving from observed surface brightness $s_{\mathrm{Ly}\alpha}$ to intrinsic 'surface luminosity' $S_{\mathrm{Ly}\alpha}$, expressed in erg s^{-1} kpc^{-2}. Furthermore, we accounted for the inevitable faint-end incompleteness of the Lyα emitter sample by tying the integration to a completeness-corrected population distribution statistic (Methods). The resulting cumulative incidence rates as functions of surface luminosity threshold are presented in Fig. 4a. Values of dn/d$z > 1$ indicate that a random line of sight passes on average through more than one emitter within a redshift interval of $\Delta z = 1$.

In Fig. 4b we compare our measured incidence rates with the statistics of atomic hydrogen detected in absorption against background quasars[11–14]. We find that emission and absorption incidence rates dn/dz_{em} and dn/dz_{abs} have a similar range of values, which we use to tentatively match surface luminosities to column densities.

Emission regions with $\log_{10}[S_{\mathrm{Ly}\alpha}$ (erg s^{-1}kpc^{-2})] $\gtrsim 38$ (for brevity we omit the units of $S_{\mathrm{Ly}\alpha}$ in the following), typically at radial distances of less than about 2″, have a dn/dz_{em} of about 0.5 per unit redshift, which is comparable to damped Lyα absorbers[11,14] (DLAs) with column densities of $\log_{10}[N(\text{H I})\ (\mathrm{cm}^{-2})] > 20.3$. At redshifts $z \lesssim 3.5$ this result broadly agrees with previous findings[20] based on long-slit spectroscopy of a much smaller sample. Our data also show that the trend of dn/dz with redshift is very similar for absorbers and emitters. It is thus plausible to identify DLAs with Lyα-emitting regions at levels of $\log_{10}(S_{\mathrm{Ly}\alpha}) > 38$, which is also approximately the limit for the detection of individual Lyα haloes[4,5]. This Lyα emission is likely to be powered by ultraviolet photons from star-forming regions and then resonantly scattered outwards[21–23], possibly enhanced by cooling radiation during the accretion of gas into dark matter haloes[24–26].

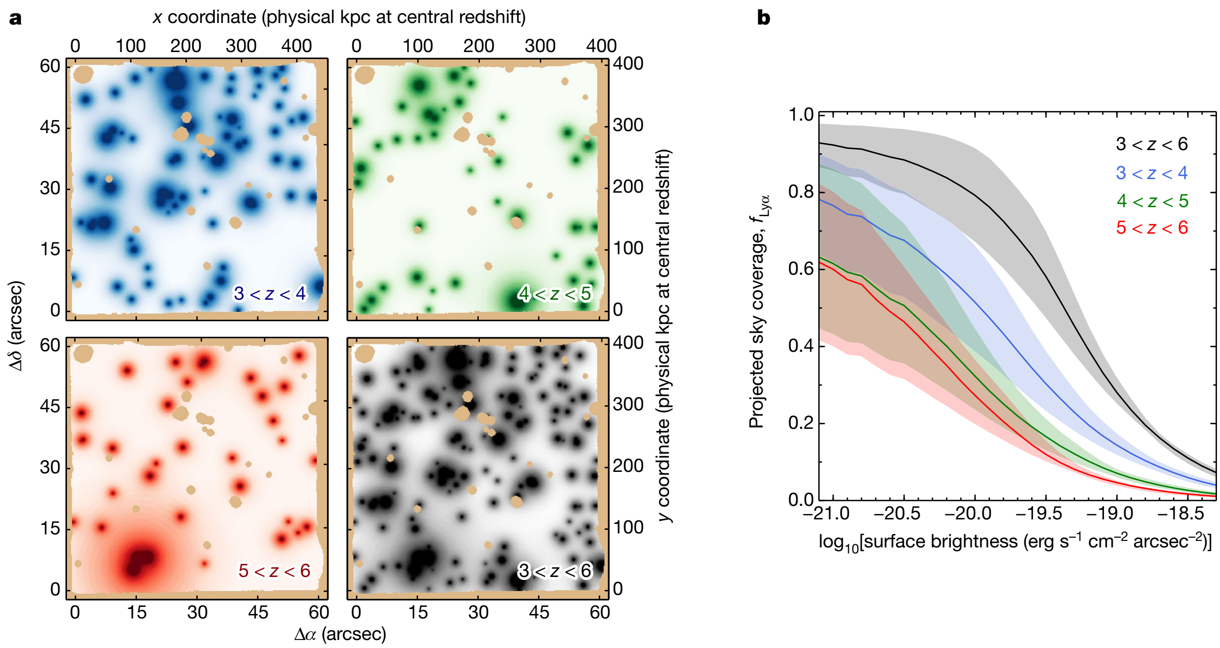


Fig. 3 | The Lyα sky coverage from median-stacked profiles. **a**, Reconstructed noise-free Lyα model images of the $1' \times 1'$ section of the HUDF, shown separately for the three disjoint redshift bins (given at bottom right of each panel, and indicated by the blue/green/red colours) and for the full redshift range (shown in black). The shading indicates surface brightness in logarithmic stretch. At the observed position of each object a source model was inserted after rescaling it to its actual Lyα flux. The light brown areas delineate the MUSE field of view and also mask bright foreground objects. α, right ascension; δ, declination. **b**, The cumulative fractional sky coverage of projected Lyα emission, as a function of limiting surface brightness. Labels and colours indicate the redshift ranges. The solid lines represent the average relations from our two MUSE pointings, while the shaded bands outline our 1σ uncertainty estimates from combining the errors of the profiles and the differences between the two fields.

Moving to lower column densities of atomic hydrogen, Fig. 4b shows that systems with $\log_{10}[N(\mathrm{H\,I})\ (\mathrm{cm}^{-2})] > 19$ (sometimes called sub-DLAs or SDLAs) and Lyα emission regions with $\log_{10}(S_{\mathrm{Ly}\alpha}) \gtrapprox 37.5$ both have dn/dz of order unity. At even lower N(H I), Lyman limit systems (LLSs) with $\log_{10}[N(\mathrm{H\,I})\ (\mathrm{cm}^{-2})] > 17$ give rise to several incidences per unit redshift[12], which can be approximately matched in emission by surface luminosities of $\log_{10}(S_{\mathrm{Ly}\alpha}) \gtrapprox 37$, a level just detectable in our median stack at $3 < z < 4$ (but requiring some extrapolation of the profiles for $z > 4$). Above column densities of about 10^{18} cm^{-2}, hydrogen becomes self-shielded to ionizing radiation and thus at least partly atomic[27], although most of the gas in this N(H I) range is still ionized. In fact nearly all atomic hydrogen at $z > 3.5$ is found[11] in absorbers with

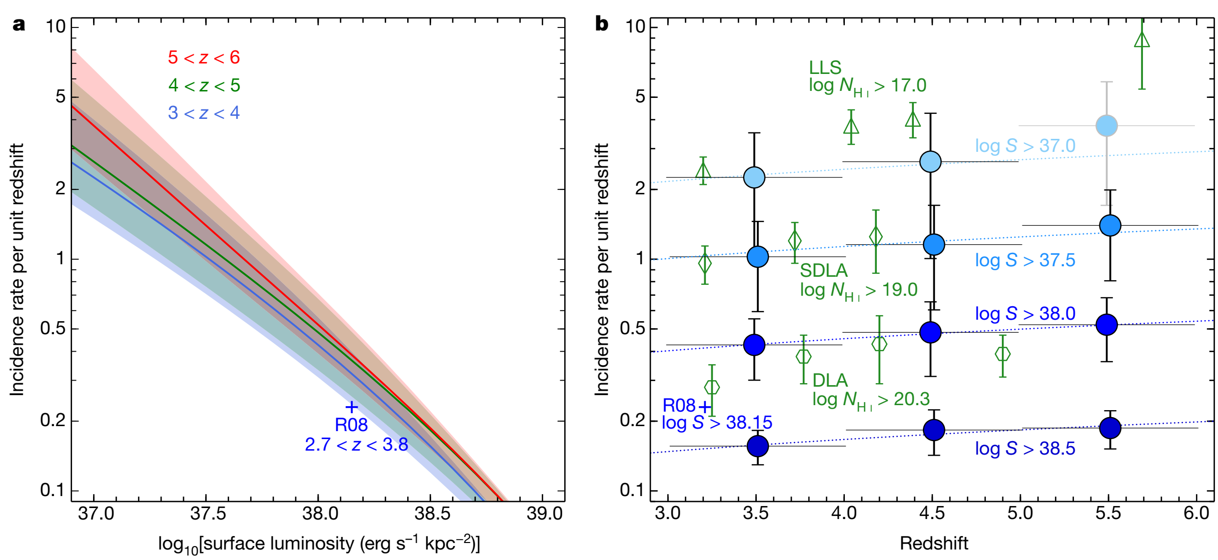


Fig. 4 | Incidence rates of Lyα emission and comparison with absorption measurements. **a**, Cumulative incidence rates of Lyα in emission, as a function of limiting surface luminosity; data are shown colour-coded for the three redshift ranges given. The shaded bands outline our estimated 1σ uncertainties. **b**, Blue filled circles show the inferred evolution of Lyα emission incidence rates with redshift, for four different thresholds in Lyα surface luminosity labelled by their log$S$ values. In order to avoid overlapping error bars (1σ as defined in Methods) the symbols are offset by ± 0.01 in redshift. The top-right datapoint is plotted in light grey to indicate that it involves extrapolation of the Lyα radial profiles and is less certain than the other points. The blue cross (R08) in both panels represents the only previous observational estimate of the Lyα emission incidence rate[20]. Green open symbols show literature values of the cumulative incidence rates of atomic hydrogen measured from quasar absorption lines, for three commonly adopted limits in column density, $N_{\mathrm{H\,I}}$ (triangles, Lyman limit systems[12], LLS; diamonds, sub-damped Lyα absorbers[13], SDLA; hexagons, damped Lyα absorbers[11,14], DLA). The thin dotted lines show the expected trend for an intrinsically non-evolving population of Lyα emitters.

$\log_{10}[N(\text{H I})\ (\text{cm}^{-2})] > 19$. The approximate equality of the incidence rates of H I absorbers with column densities $\log_{10}[N(\text{H I})\ (\text{cm}^{-2})] > 19$ and Lyα-emitting regions with surface luminosities $\log_{10}(S_{\text{Ly}\alpha}) \gtrsim 37.5$ therefore suggests that in these regions we are observing the faint glow of ubiquitous circumgalactic atomic hydrogen. This result is robust with respect to modest deviations from the assumed azimuthal symmetry in the spatial distribution of the emitting gas (Methods). The main systematic uncertainty, not included in our error bars, lies in the fact that our sample is selected by its Lyα emission. We estimate that we captured roughly 50% of the galaxies at these redshifts (Methods), implying that the incidence rates of circumgalactic absorbers should be reduced by this factor before comparing with the emission incidence rates. These uncertainties do not affect our conclusion that most atomic hydrogen at redshifts 3 to 6 has now also been detected in emission.

Our results suggest that $\mathrm{d}n/\mathrm{d}z_{\text{em}}$ increases mildly with redshift. In Fig. 4b we compare our measurements with the expected redshift dependence for an intrinsically non-evolving population of emitters. Such a population can be described by a constant incidence rate $\mathrm{d}n/\mathrm{d}X_{\text{em}}$ per comoving path length X along the line of sight, a quantity commonly used in quasar absorption line studies. Within the redshift range considered here, a constant $\mathrm{d}n/\mathrm{d}X(z)$ translates into a $\mathrm{d}n/\mathrm{d}z(z)$ similar to the trend suggested by our data points. Because the Lyα luminosity function also shows very little evolution within the redshift range of our sample[28], we conclude that the properties of circumgalactic Lyα emission do not change much between redshifts 6 and 3. Nevertheless, given the error bars the data are also compatible with a constant $\mathrm{d}n/\mathrm{d}z_{\text{em}}(z)$, corresponding to a modest decrease of incidence rates per comoving path length.

What powers this low-level Lyα emission that we tentatively identify as originating from circumgalactic high-column-density absorbers? At large galactocentric distances, Lyα fluorescence of optically thick gas excited by the cosmic ultraviolet background (UVB) becomes a viable possibility[3]. The expected fluorescent Lyα surface brightness of an LLS at $z = 3$ has been calculated[29] and recently updated[30] for $z = 3.5$ as $s_{\text{Ly}\alpha,\text{UVB}} \approx 1.1 \times 10^{-20}\ \text{erg s}^{-1}\ \text{cm}^{-2}\ \text{arcsec}^{-2}$, which is just about within the sensitivity range of our stacked data and consistent with the marginal signal at large radii in our lowest redshift bin. This suggests that at least some of the extremely faint Lyα emission detected by MUSE may be due to this omnipresent glow, opening a window to an important but previously invisible component of the cosmic matter distribution.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0564-6.

Received: 15 March 2018; Accepted: 1 August 2018;
Published online 1 October 2018.

1. Partridge, R. B. & Peebles, P. J. E. Are young galaxies visible? *Astrophys. J.* **147**, 868–886 (1967).
2. Hogan, C. J. & Weymann, R. J. Lyman-alpha emission from the Lyman-alpha forest. *Mon. Not. R. Astron. Soc.* **225**, 1P–5P (1987).
3. Gould, A. & Weinberg, D. H. Imaging the forest of Lyman limit systems. *Astrophys. J.* **468**, 462–468 (1996).
4. Wisotzki, L. et al. Extended Lyman α haloes around individual high-redshift galaxies revealed by MUSE. *Astron. Astrophys.* **587**, A98 (2016).
5. Leclercq, F. et al. The MUSE Hubble Ultra Deep Field Survey. VIII. Extended Lyman-alpha haloes around high-*z* star-forming galaxies. *Astron. Astrophys.* **608**, A8 (2017).
6. Hayes, M., Scarlata, C. & Siana, B. Central powering of the largest Lyman-α nebula is revealed by polarized radiation. *Nature* **476**, 304–307 (2011).
7. Cantalupo, S., Arrigoni-Battaia, F., Prochaska, J. X., Hennawi, J. F. & Madau, P. A cosmic web filament revealed in Lyman-α emission around a luminous high-redshift quasar. *Nature* **506**, 63–66 (2014).
8. Hennawi, J. F., Prochaska, J. X., Cantalupo, S. & Arrigoni-Battaia, F. Quasar quartet embedded in giant nebula reveals rare massive structure in distant universe. *Science* **348**, 779–783 (2015).
9. Steidel, C. C. et al. Diffuse Lyα emitting halos: a generic property of high-redshift star-forming galaxies. *Astrophys. J.* **736**, 160–177 (2011).
10. Momose, R. et al. Diffuse Lyα haloes around galaxies at $z = 2.2$–6.6: implications for galaxy formation and cosmic reionization. *Mon. Not. R. Astron. Soc.* **442**, 110–120 (2014).
11. Péroux, C., McMahon, R. G., Storrie-Lombardi, L. J. & Irwin, M. J. The evolution of Ω_{HI} and the epoch of formation of damped Lyman α absorbers. *Mon. Not. R. Astron. Soc.* **346**, 1103–1115 (2003).
12. Songaila, A. & Cowie, L. L. The evolution of Lyman limit absorption systems to redshift six. *Astrophys. J.* **721**, 1448–1466 (2010).
13. Zafar, T. et al. The ESO UVES advanced data products quasar sample. *Astron. Astrophys.* **556**, A141 (2013).
14. Crighton, N. H. M. et al. The neutral hydrogen cosmological mass density at $z = 5$. *Mon. Not. R. Astron. Soc.* **452**, 217–234 (2015).
15. Bacon, R. et al. The MUSE second-generation VLT instrument. *Proc. SPIE* **7735**, 773508 (2010).
16. Bacon, R. et al. MUSE commissioning. *The Messenger* **157**, 13–16 (European Southern Observatory, Garching, 2014).
17. Bacon, R. et al. The MUSE 3D view of the Hubble Deep Field South. *Astron. Astrophys.* **575**, A75 (2015).
18. Casertano, S. et al. WFPC2 observations of the Hubble Deep Field South. *Astron. J.* **120**, 2747–2824 (2000).
19. Beckwith, S. V. W. et al. The Hubble Ultra Deep Field. *Astron. J.* **132**, 1729–1755 (2006).
20. Rauch, M. et al. A population of faint extended line emitters and the host galaxies of optically thick QSO absorption systems. *Astrophys. J.* **681**, 856–880 (2008).
21. Laursen, P. & Sommer-Larsen, J. Lyα resonant scattering in young galaxies: predictions from cosmological simulations. *Astrophys. J.* **657**, L69–L72 (2007).
22. Barnes, L. A. & Haehnelt, M. G. Faint extended Lyα emission due to star formation at the centre of high column density QSO absorption systems. *Mon. Not. R. Astron. Soc.* **403**, 870–885 (2010).
23. Verhamme, A. et al. Lyman-α emission properties of simulated galaxies: interstellar medium structure and inclination effects. *Astron. Astrophys.* **546**, A111 (2012).
24. Haiman, Z., Spaans, M. & Quataert, E. Lyα cooling radiation from high-redshift halos. *Astrophys. J.* **537**, L5–L8 (2000).
25. Fardal, M. A. et al. Cooling radiation and the Lyα luminosity of forming galaxies. *Astrophys. J.* **562**, 605–617 (2001).
26. Furlanetto, S. R., Schaye, J., Springel, V. & Hernquist, L. Lyα emission from structure formation. *Astrophys. J.* **622**, 7–27 (2005).
27. Altay, G., Theuns, T., Schaye, J., Crighton, N. H. M. & Dalla Vecchia, C. Through thick and thin – H I absorption in cosmological simulations. *Astrophys. J.* **737**, L37 (2011).
28. Drake, A. B. et al. The MUSE Hubble Ultra Deep Field Survey. VI. The faint-end of the Lyα luminosity function at $2.91 < z < 6.64$ and implications for reionisation. *Astron. Astrophys.* **608**, A6 (2017).
29. Cantalupo, S., Porciani, C., Lilly, S. J. & Miniati, F. Fluorescent Lyα emission from the high-redshift intergalactic medium. *Astrophys. J.* **628**, 61–75 (2005).
30. Gallego, S. G. et al. Stacking the cosmic web in fluorescent Lyα emission with MUSE. *Mon. Not. R. Astron. Soc.* **475**, 3854–3869 (2018).

Acknowledgements We thank the ESO staff for the support that made these observations possible. L.W., J.K., R.S. and T.U. acknowledge support by the Competitive Fund of the Leibniz Association through grants SAW-2013-AIP-4 and SAW-2015-AIP-2. R.B., H.I., F.L. and M.A. are supported by the ERC advanced grant 339659-MUSICOS. J.B. acknowledges support by FCT grants UID/FIS/04434/2013 and IF/01654/2014/CP1215/CT0003 and by FEDER through COMPETE2020 (POCI-01-0145-FEDER-007672). J.R. acknowledges support from the ERC starting grant 336736-CALENDS. P.M.W. received support through BMBF Verbundforschung, grant 05A17BAA. T.C., N.B. and B.G. acknowledge support by ANR FOGHAR (ANR-13-BS05-0010-02). T.C. and N.B. were also supported by OCEVU Labex (ANR-11-LABX-0060), and by the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'avenir" French government programme. A.M.I. acknowledges support from MINECO through project AYA2015-68217-P. S.C. acknowledges support from Swiss National Science Foundation grant PP00P2_163824.

Reviewer information *Nature* thanks G. Micheva and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.W. conceived the project. R.B. led the data acquisition and data reduction. R.B., J.B., E.C.H., H.I., J.K., K.B.S., T.U. and L.W. developed and performed the sample selection. L.W. analysed the data, with input from R.B., J.B. and P.M.W. S.C., P.R., J.S., M.S. and L.W. worked on the interpretation of the results. L.W. wrote the manuscript and produced the figures, with K.B.S. contributing to their design. All coauthors provided critical feedback to the text and helped shape the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0564-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0564-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

MUSE observations. MUSE is an integral field spectrograph mounted on Unit Telescope 4 of the ESO Very Large Telescope. In its Wide Field Mode it offers a $1' \times 1'$ field of view at a spatial sampling of $0.2'' \times 0.2''$, producing a datacube with 90,000 spatial pixels. Each spatial pixel contains a 475–935 nm spectrum with $\sim$0.25 nm spectral resolution. The first of the two deep field observations used here was obtained in 2014, where we integrated for a total of 27 h on a single pointing in the Hubble Deep Field South[18] (HDFS). The data reduction and construction of the first redshift catalogue in this field are summarized elsewhere[17]. The second deep MUSE exposure was obtained between 2014 and 2016 as part of a greater effort to perform a contiguous spectroscopic mapping of the Hubble Ultra-Deep Field[19] (HUDF). These observations resulted in a $3' \times 3'$ mosaic with a mean integration time of 10 h, on top of which 21 h of additional exposure time were dedicated to a single MUSE pointing inside the HUDF. The characteristics of the MUSE-HUDF data set and the reduction process are described elsewhere[31]. Here we use only the two ultra-deep 1 arcmin^2 MUSE pointings, which for simplicity we refer to as HDFS and HUDF, respectively. The average spatial resolution (FWHM of the best-fitting Moffat[32] function) in the combined coadded datacube, evaluated at 700 nm, is $0.66''$ for the HDFS and $0.63''$ for the HUDF.

The sample of Lyα emitters. Here we focus on galaxies marked by their Lyα emission (Lyα emitters, LAEs). Given the MUSE spectral range of 475–935 nm, LAEs can be detected over a redshift interval of $2.92 < z < 6.64$. In order to construct a homogeneous Lyα-selected sample, we ran our dedicated software LSDCat[33] (Line Source Detection and Cataloguing) with a signal-to-noise ratio threshold of 6 on both datacubes to produce a list of emission line objects, which we then inspected visually to assign redshifts. This resulted in a sample of 128 LAEs in the HDFS and 161 LAEs in the HUDF, respectively. Because of the strictly Lyα-based selection, these samples are not mere subsets of our previously published catalogues[17,34] but also contain a few additional LAEs. The spatial distribution of the objects is visualized in Extended Data Fig. 1. For each LAE, the LSDCat software measures the spatial and spectral centroids and the integrated Lyα fluxes, quantities used in this study. Redshifts were assigned from the measured centroids of the Lyα emission line. While these centroids are known to be shifted with respect to the systemic redshifts by up to a few hundred km s^{-1}, accurate knowledge of the latter is not required for the present study (and was not available for our sample). Because of the decrease in instrument sensitivity close to the low and high wavelength cutoffs, and because of the crowding of OH night-sky emission lines towards the reddest wavelengths, we limited our sample to the redshift range $3 < z < 6$, which conveniently allowed us to define three broad redshift bins of $\Delta z = 1$ each. The final sample encompasses 119 LAEs in the HDFS and 151 LAEs in the HUDF, respectively. We provide the full sample as a data table (Supplementary Data) in machine-readable format containing the relevant quantities for this study: positions, redshifts, integrated Lyα fluxes, and flags indicating which objects were used for the direct projection and median stacking subsets. The newly found LAEs will also be included in a forthcoming public catalogue update for these fields.

Extraction of narrowband images. Each LAE enters into the current investigation as a pseudo-narrowband (NB) Lyα image, extracted from the datacube at the location of the three-dimensional Lyα centroid coordinates provided by LSDCat. These images were constructed in the following way: we first extracted a provisional Lyα spectrum from the continuum-subtracted datacube by summing over an unweighted circular aperture of radius $0.6''$. We then modelled the Lyα emission line profile as a Gaussian, which provided an improved line centroid as well as an approximate line width. Using this Gaussian approximation as spectral template, we performed a weighted summation of spectral layers of the datacube into a single NB image. While Lyα lines usually show some deviations from a single Gaussian, these deviations have negligible effect on the extraction results, except for cases of secondary line peaks ('blue bumps') which are not captured by our narrowband images. The maximum Gaussian FWHM for the extraction was set to 500 km s^{-1} in order to limit the noise. Compared to an unweighted summation over a given bandwidth this scheme provides a better signal-to-noise ratio, and the fractional weights ensure that the bandwidth always matches the actual line width, which matters especially for relatively narrow lines with FWHM $\lessapprox$ 200 km s^{-1}. We verified that for broader lines the results are very similar to an unweighted summation. The blank-sky noise level in these NB images varies substantially, depending on the spectral bandwidth and on the wavelength of the Lyα line. A typical value of the pixel-to-pixel r.m.s. in regions with no detected emission is 5×10^{-19} erg s^{-1} cm^{-2} pixel^{-1}, corresponding to a 1σ surface brightness limit of 10^{-19} erg s^{-1} cm^{-2} arcsec^{-2} when averaged over an aperture of $1''$. This limit varies by a factor of $\sim$2 between different objects.

The Lyα sky coverage from direct projection. We obtained a projected Lyα view of each field by coadding all extracted NB images, maintaining the position of each object in the plane of the sky. In order to reduce the noise in the coadded image, we introduced a truncation radius of $6''$ around the centroid of each LAE beyond which the NB data were set to zero for the coadding procedure. This was motivated by the fact that beyond this radius we find generally no individually detectable Lyα emission around our objects. The projection therefore accounts only for the circumgalactic Lyα emission around detected sources. Any putative extended Lyα emission at the same redshift as a detected object, but outside the truncation radius, is ignored in the procedure. At the same time the truncation ensured that same-redshift pairs entered only once into the coadded image; if necessary we constructed additional masks by hand to ensure that this was always the case. Masks were also applied in a few cases of contamination by foreground emission lines or by continuum subtraction residuals from bright foreground objects such as stars. Finally, we removed 33 objects where Lyα falls close to a bright sky line causing significant sky-subtraction residuals in the narrowband images. Altogether the amount of budgeted Lyα flux in this approach should be seen as a strict lower limit to the true emission in the field.

With these provisions, we estimated the cumulative fractional sky coverage $f_{\mathrm{Ly}\alpha}$ of Lyα emission brighter than $s_{\mathrm{Ly}\alpha}$, documented in Extended Data Fig. 2. Without spatial filtering the surface brightness is given for a single pixel of only $0.2'' \times 0.2''$, resulting in extremely noisy images. We therefore filtered the images with a Gaussian of FWHM = 7 pixels ($1.4''$) which provided a good compromise between noise suppression, enforcing large-scale spatial coherence, and ensuring that flux redistribution from the central pixels into the outer parts can be neglected. The maximally reachable covering fraction is limited by random fluctuations due to noise, as can be seen very clearly in Extended Data Fig. 2: $f_{\mathrm{Ly}\alpha}$ measured from the unfiltered data converges to considerably lower values than $f_{\mathrm{Ly}\alpha}$ measured from the filtered data—which are of course also affected, but to a lesser degree. We have not attempted to push this approach any further, as for the main results of this paper we employed the stacking-and-insertion modelling approach described below. We did, however, perform a retrospective consistency check between the direct projection and the stacking approach, as follows: we took the idealized noise-free reconstructed model images obtained from the median-stacking analysis (the bottom-right image in Fig. 3a in the case of the HUDF), degraded them by adding realistic noise, and after spatial filtering we measured $f_{\mathrm{Ly}\alpha}$ in the same way as in the real data. For the noise model we filled empty datacubes with normally distributed random numbers scaled to the effective noise in the actual data. From these noise-only cubes we extracted NB images with the same prescriptions as for the real LAEs, using the same spectral bandpasses and spatial masks, and coadded them to provide a random noise realization of the projected Lyα image, which we then added to the stacking-based model image. The grey curves in Extended Data Fig. 2 show that these very different approaches to measuring $f_{\mathrm{Ly}\alpha}$ produce remarkably similar results, considering that the direct projection approach does not involve azimuthal averaging and is based on an incomplete sample without any completeness corrections. The only noteworthy discrepancy between the thick black and the thick grey lines in Extended Data Fig. 2 occurs around $\log_{10}(s_{\mathrm{Ly}\alpha}) \approx -18.5$, mainly caused by the median-taking in the stacking process which removes the largest and most extended Lyα emitters. At these relatively high surface brightnesses, direct projection actually delivers a more realistic estimate of $f_{\mathrm{Ly}\alpha}$ than the stacking approach.

Stacking analysis. We excised $20'' \times 20''$ MUSE-narrowband subimages centred on each source and put these into image stacks, separately for the three redshift intervals $3 < z < 4$, $4 < z < 5$ and $5 < z < 6$. Before the analysis, the data were subjected to a rigorous visual screening. In this step, 76 objects were removed from the stacks because their NB images were disturbed by sky subtraction residuals or residual emission from unrelated objects. 194 LAEs remained in the combined stacking sample. Bright foreground objects and the edges of the field of view were masked. We also ensured that when the NB image contained multiple objects at the same redshift, each spatial pixel contributed to the stack only once. We identified 13 double sources, 2 triples and one quadruple for which the subimage had thus to be divided up, by drawing a line through the image and assigning each pixel to only one object, or by using only pixels inside of a certain radius, typically $6''$. Finally, each stack was collapsed into a single image by computing the pixel-by-pixel median of all unmasked input image pixels. We chose the median instead of the mean to avoid the possibility of faint undetected companions enhancing the signal in the outskirts, and to make the stacked images robust against faint artefacts escaping from the visual screening. For each collapsed stack we also obtained a weight image containing in each pixel the number of input images that contributed to it after masking; these weight images are shown in Extended Data Fig. 3.

Profile extraction and error estimation. To detect the low-surface-brightness Lyα emission in the outer regions of our haloes, we extracted azimuthally averaged radial profiles from the median stacks (Fig. 2c), measured by averaging all pixels within each of a set of concentric annuli (Fig. 2b) defined as follows. Let r_i denote the outer radius of annulus i with $i = 1, \ldots, 11$. For $i \leq 5$ we adopted constant annular widths, $r_i = i \times 0.2''$ (1 MUSE spatial pixel), for the outer annuli $i \geq 6$ we constructed a progression of increasing widths with the recursion formula $r_i = r_{i-1} + 10^{1.47(i-5)/7} \times 0.2''$. Thus, the last annulus has an outer radius $r_{11} = 9.97''$, just fitting into our $20'' \times 20''$ images and combining 4,660 MUSE spatial pixels into a single mean surface brightness measurement.

We estimated the uncertainties of these surface brightness profiles in two ways, by formal error propagation and empirically from empty regions in the data. The formal errors originate in the pixel variances of the MUSE datacubes, corrected for resampling effects[31], propagated first into the NB images of individual objects and then into the median-combined stacks. For the latter step we used the property of the median that its variance is approximately $\pi/2$ times the variance of the mean. For the empirical noise calibration we shifted the whole LAE sample in wavelength while maintaining the spatial positions of all objects, thus defining empty regions by applying small redshift offsets. Offsetting in increments of ±5 MUSE spectral pixels or ±6.25 Å yielded 40 complete sets of as many empty regions as LAEs, which were then subjected to the same NB image extraction and median stacking procedure as the Lyα data. We estimated the noise from the dispersion of the extracted radial profiles between these 40 sets. The outcome of this experiment is presented in Extended Data Fig. 4. This figure thus provides the significance limits for the detection of very low-surface-brightness emission in our stacked data.

Although the propagated errors can be easily calculated independently of the object content of the datacubes, they do not account for systematics in the background subtraction. On the other hand, the empirically determined errors automatically include such systematics, but are subject to contamination of the supposedly empty regions by unrelated sources and by sky subtraction residuals, and hence they are likely to overestimate the true errors. In our experiment the median ratios between empirical and propagated errors for the outer profile regions ($r > 1''$) were (1.02, 1.33, 1.52) for the three redshift ranges. For this paper we conservatively adopted only the (larger) empirically determined errors for the profiles, and only these are shown in Fig. 2c and Extended Data Fig. 4. The empty regions experiment also revealed that the mean of the empty regions tends to become slightly negative at small radii $r < 2''$, implying that we may actually underestimate the surface brightnesses in the inner regions by a very small amount. Because this effect is at the ~1% level relative to the measured signal at these radii, we decided to neglect it.

Surface brightness modelling. We used GALFIT[35] to model the median-stacked images with smooth 2-dimensional Lyα surface brightness distributions. While in previous work[4,5] we favoured a double exponential model with a compact core and an extended halo, that model was driven mainly by the high signal-to-noise ratio and high surface brightness regions within $\lesssim$10 kpc. In the present study the emphasis is on the outer regions, where Fig. 2 suggests that the surface brightness distribution of the halo may show some flattening relative to a single exponential. Here we adopted a model consisting of a central point source plus a circular Sersic[36] function, convolved with the point-spread function (PSF) at the relevant wavelengths, which provided good fits to all median-stacked images. To quantify the uncertainties in the fitted profiles we used the formal error estimates of GALFIT, but then increased these until the uncertainties of the fits were consistent with the empirically calibrated error bars of the directly extracted azimuthal profiles. These uncertainties are displayed as the shaded regions around the fitted profiles in Fig. 2. Extended Data Table 1 provides the numerical values of the fit parameters and their uncertainty estimates.

We make the central assumption that the shapes of the Lyα haloes of LAEs do not, in the statistical average, depend on their luminosities. While in ref. [5] we found no evidence for such a dependence, most of those objects have Lyα-luminosities $L_{\mathrm{Ly}\alpha} > 10^{42}$ erg s^{-1}, whereas our current sample has a median $\log_{10}[L_{\mathrm{Ly}\alpha}\,(\mathrm{erg\,s^{-1}})]$ of only 41.7. Here we use the median stacks to probe deeper into the validity of the above assumption. In Extended Data Fig. 5 we present a comparison between the radial profiles extracted from the median stack of the full sample and of a subset with Lyα luminosities $L_{\mathrm{Ly}\alpha} > 10^{42}$ erg s^{-1}. There are (18, 13, 6) objects at z = (3–4, 4–5, 5–6) meeting this criterion. The median-stack profiles of the luminous subsets resemble those of the full sample remarkably well. This comparison supports the validity of our self-similarity approximation across the luminosity range of our sample. We plan to revisit this point and related aspects in a follow-up study of Lyα halo profile shapes in stacked MUSE data.

We used the analytic profile fits as templates to construct idealized representations of all LAEs in the two fields, including those objects previously removed from the stacking subsample. All model LAEs at a given redshift range were assigned to have the same spatial surface brightness distribution, but rescaled to the actually observed Lyα fluxes of each real object. Furthermore, the GALFIT models were approximately corrected for PSF blurring by using a delta function (in fact a very narrow Gaussian) as PSF when reconstructing the two-dimensional model templates. Since for each object the template was rescaled to match the measured Lyα flux, this implies that the modelling of the brightest LAEs involved a certain degree of extrapolation of the profiles. We demonstrate below that the contribution of extrapolated emission to the incidence rates is small (see also Extended Data Fig. 7).

Determination of incidence rates. We estimated the Lyα emission incidence rates directly from the LAE samples as the sum of circular cross-sections $\pi r_{\mathrm{iso},i}^2$ where $r_{\mathrm{iso},i}(s_{\mathrm{Ly}\alpha})$ is the isophotal extent of the model of object i at a given surface brightness level. The incidence rate is then

$$\frac{\mathrm{d}n}{\mathrm{d}z}(S_{\mathrm{Ly}\alpha}) \approx \frac{1}{A_{\mathrm{FoV}}}\frac{n_{\mathrm{obj}}}{\sum \Delta z_i}\sum_{i=1}^{n_{\mathrm{obj}}} \pi r_{\mathrm{iso},i}^2(S_{\mathrm{Ly}\alpha}) \quad (1)$$

where A_{FoV} is the area of the field of view, Δz_i is the redshift path length over which object i would be part of the flux-limited sample, and where the summation is carried out over all objects in the redshift range. The normalization of dn/dz to a quantity per unit redshift was in our case conveniently provided by the widths of our adopted redshift intervals.

Since $s_{\mathrm{Ly}\alpha}$ decreases with increasing redshift as $(1+z)^4$, we decided to move to distance-independent surface luminosities $S_{\mathrm{Ly}\alpha}$. While this quantity is not much used in the literature, we prefer working with intrinsic object properties over rescaling observed quantities to some fiducial reference redshift. The transformation is $\log_{10}(S_{\mathrm{Ly}\alpha}) = \log_{10}(s_{\mathrm{Ly}\alpha}) + 4\log_{10}(1+z) + 54.71$ when both s and S are given in cgs units. The right-hand ordinate of Extended Data Fig. 5 provides a quick-look visual calibration of the conversion.

Correction for sample incompleteness. Our LAE sample certainly suffers from incompleteness close to the flux limit, with faint objects getting selected only if their Lyα emission is sufficiently point-like. Selection effects for the detection of LAEs in the MUSE-HUDF survey have been investigated in detail[28], and it was shown that the transition from 80% to 20% detection probability extends over ~0.5 dex in line flux. The Lyα incidence rates calculated from equation (1) are therefore biased low, missing the contributions from undetected but presumably existing objects. In order to correct for this incompleteness, we replaced the summation over the observed sample by an integration over the full survey volume, assuming that the intrinsic distribution of Lyα luminosities follows the luminosity function determined in ref. [28]. Since the self-similarity approximation of the extended Lyα emission implies a unique relation between the total flux $F_{\mathrm{Ly}\alpha}$ of an object and its isophotal radius, $r_{\mathrm{iso}} = r_{\mathrm{iso}}(F_{\mathrm{Ly}\alpha}, s_{\mathrm{Ly}\alpha})$, we can predict the emission cross-sections from only the Lyα luminosities. The total incidence rate dn/dz for a given redshift range (z_1, z_2) follows as

$$\frac{\mathrm{d}n}{\mathrm{d}z}(s_{\mathrm{Ly}\alpha}) = \frac{1}{(z_2-z_1)A_{\mathrm{FoV}}}\int_{z_1}^{z_2}\int_{\ell_{\min}}^{\ell_{\max}} \pi r_{\mathrm{iso}}^2[S_{\mathrm{Ly}\alpha}, F_{\mathrm{Ly}\alpha}(\ell,z)]\phi(\ell,z)\,\mathrm{d}\ell\,\mathrm{d}V_{\mathrm{c}}(z) \quad (2)$$

where $\mathrm{d}V_{\mathrm{c}}(z)$ is the differential comoving volume element at redshift z, $\ell \equiv \log_{10}(L_{\mathrm{Ly}\alpha}) = \log_{10}(4\pi d_{\mathrm{L}}^2 F_{\mathrm{Ly}\alpha})$ (in units of erg s^{-1}, with the cosmological luminosity distance d_{L}), and $\phi(\ell,z)\mathrm{d}\ell$ is the differential Lyα luminosity function at redshift z. We parameterized $\phi(\ell)$ as a redshift-independent Schechter function with $\ell^* = 42.59$, $\phi^* = 2.138 \times 10^{-3}$ Mpc^{-3}, and $\alpha = -1.93$, which is a good overall fit to the completeness-corrected LAE sample[28]. While the upper luminosity integration limit $\ell_{\max}$ does not matter much as $\phi(\ell)$ approaches zero very quickly for increasing ℓ, choosing a value for the lower integration limit $\ell_{\min}$ is less straightforward: Although faint LAEs have small isophotal radii, they are also numerous and thus contribute non-negligibly to the integrated cross-section. The integral converges only for $\ell_{\min} \lessapprox 40.5$, but at the expense of including large numbers of hypothetical ultrafaint LAEs into the budget that are well below the current detection limits. As our 'best guess' we adopted $\ell_{\min} = 41.0$, which is slightly brighter than the faintest detected LAEs in our actual sample. Extended Data Fig. 6 provides a synopsis of the different approaches to estimate dn/dz from our data. These plots also show that the magnitude of the completeness correction is by far the dominant source of uncertainty for dn/dz. We therefore adopted as the lowest reasonable limit the values of dn/dz without any completeness correction (that is, from equation (1)) obtained from the somewhat 'emptier' HDFS. As an upper limit we took the asymptotic result from integrating equation (2) with $\ell_{\min} = 40.0$. For the presentation in Fig. 4 we interpreted these lower and upper bounds as $\pm 2\sigma$ limits, but plotted only the 1σ error envelopes in accordance with the usual conventions.

Discussion of systematic errors. We first consider to what extent the integrated values of dn/dz depend on extrapolations of the rescaled Lyα profiles beyond the radial range over which they were constructed. To address this question, we constructed the cumulative distribution of the contributions to the total incidence rate sum or integral as a function of their isophotal radii. The results of the completeness-corrected integration of dn/dz (equation (2)) with $\ell_{\min} = 41.0$ (our 'best guess' approach) are shown in Extended Data Fig. 7. These plots demonstrate that at all redshifts and for all levels of $S_{\mathrm{Ly}\alpha}$ except the lowest, more than 80% of the total incidence rates is contributed by emission from $r_{\mathrm{iso}} < 5''$. Only for $\log_{10}[S_{\mathrm{Ly}\alpha,\mathrm{lim}}\,(\mathrm{erg\,s^{-1}\,kpc^{-2}})] = 37$ and $5 < z < 6$ there is a substantial extrapolated contribution, which is why we consider this point as uncertain and plotted it in light grey in Fig. 4b.

A strong assumption made in our analysis is the azimuthal symmetry of the Lyα emission, which is certainly not strictly correct. We now quantify the biases arising from the circularization of a non-axisymmetric signal through median stacking. Extended Data Fig. 8a–c shows how an elliptical two-dimensional surface brightness distribution and the corresponding isophotal cross-sections are

modified when the cross-sections are estimated from an azimuthally averaged profile. The circularized profile is broadened and the cross-sections at given surface brightnesses are overestimated, but the effect is quite small. More relevant for our analysis, Extended Data Fig. 8d–f demonstrates that median stacking of randomly oriented elongated objects delivers isophotal cross-sections that are systematically smaller than the true values. While the above calculations are based on a rather simple source model, the conclusion can be qualitatively generalized to other non-axisymmetric surface brightness distributions. The median stacking ensures that the derived emission cross-sections, and consequently the inferred Lyα sky coverage and incidence rates, are rather under- than overestimated.

While the small scale distribution of Lyα emission remains unknown at these surface brightness levels, we have some idea of the projected covering fraction $f_{\rm H\,I}$ of neutral hydrogen close to galaxies. Ref. [37] used the cosmological EAGLE simulation[38] to show that $f_{\rm H\,I}$ depends on many parameters: distance to galaxy centre, column density, redshift, halo mass, environment. Measurements[39] of the H I absorption line close to galaxies at $z \approx 2.5$ give $f_{\rm H\,I} = 0.3 \pm 0.14$ for Lyman limit systems within one virial radius ($r_{\rm vir}$); there are so far no good measurements for $z > 3$. Simulations predict that $f_{\rm H\,I}$ increases rapidly towards higher redshift[37], and we expect $f_{\rm H\,I} \approx 0.4$–0.8 for $r < r_{\rm vir}$ at the redshifts of our sample. The virial radii of our LAEs are unknown, but are predicted to be around 30 kpc or less[40]. Consulting again Extended Data Fig. 7, we see that except for $\log_{10}[S_{\rm Ly\alpha,lim}\,({\rm erg\,s^{-1}\,kpc^{-2}})] = 37$ and $5 < z < 6$, more than 80% of the integrated Lyα emission incidence rates come from radii less than 30 kpc, that is, from within one virial radius. Unless the Lyα-emitting gas has a very different spatial distribution from the general circumgalactic H I, the covering fractions $f_{\rm H\,I}$ are expected to be sufficiently close to unity that the systematic errors from any non-axisymmetry of our derived incidence rates should be small.

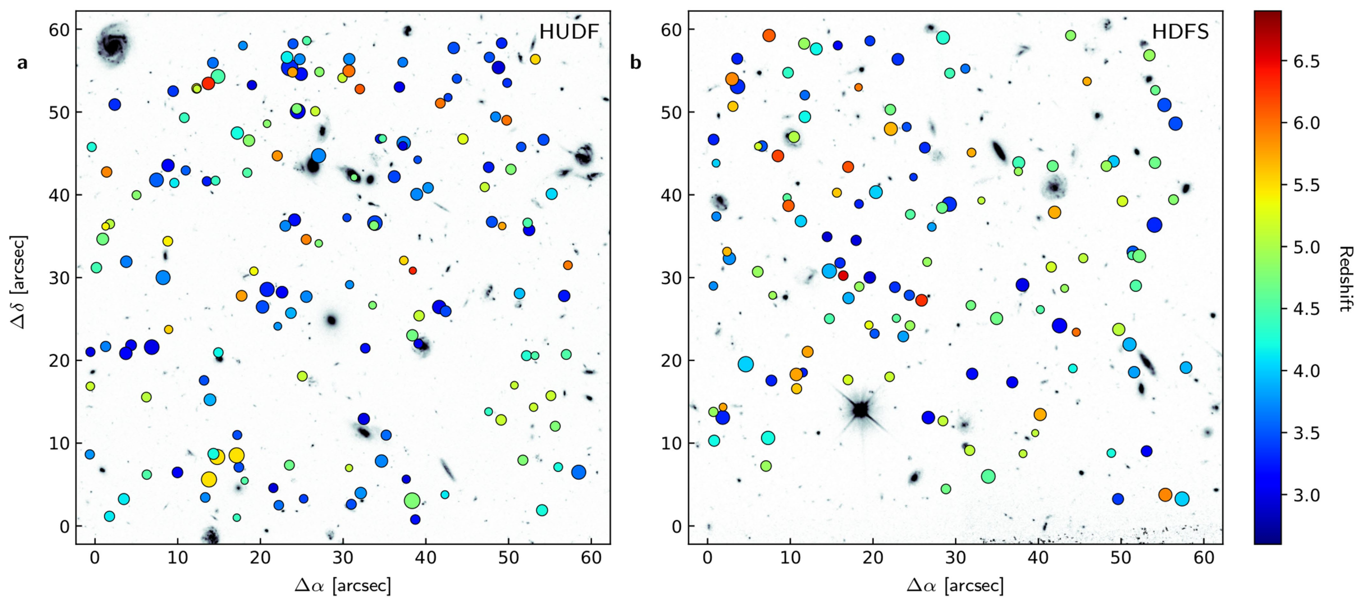
A rather different systematic error arises from the limitation of our sample to galaxies selected by their Lyα emission. If not all galaxies are LAEs, then there is circumgalactic H I gas contributing to high-column-density absorption systems which is not included in our budget of extended Lyα emission. The fraction of galaxies at $z > 3$ showing detectable Lyα emission depends strongly on the selection criteria. While only some 10%–20% of continuum-bright galaxies at these redshifts are[41] strong LAEs with Lyα rest-frame equivalent widths (EW) greater than 50 Å, this fraction probably increases to $\sim$50% if weaker emitters are also included[9,42]. There are indications that the fraction of strong emitters may even be considerably larger than 50% for very low luminosity galaxies and/or higher redshifts[41,43]. It seems thus plausible to estimate that our Lyα selection captured roughly half of all galaxies at these redshifts. If the other 50% have a circumgalactic medium similar to that of the LAEs except for the lack of Lyα emission (this is a very uncertain assumption), the incidence rates of the circumgalactic absorbers in Fig. 4 should be shifted downwards by $\sim$0.3 dex to account only for the LAE fraction. Interpolating between the measured values, a surface luminosity level of $\log_{10}[S_{\rm Ly\alpha}\,({\rm erg\,s^{-1}\,kpc^{-2}})] \approx 37.5$ would then have roughly the same incidence rate as absorbers with $\log_{10}[N({\rm H\,I})\,({\rm cm^{-2}})] \approx 18$, less than the limit for SDLAs but still optically thick to Lyman continuum radiation. On the other hand, there may also be non-LAE galaxies with still undetected faint Lyα haloes, similar to those found in ref. [9], which would increase the Lyα incidence rates even further.

Code availability. This study was carried out with several small custom routines, mostly in Python. While the analysis steps are described in the paper, individual pieces of code can be provided upon request.

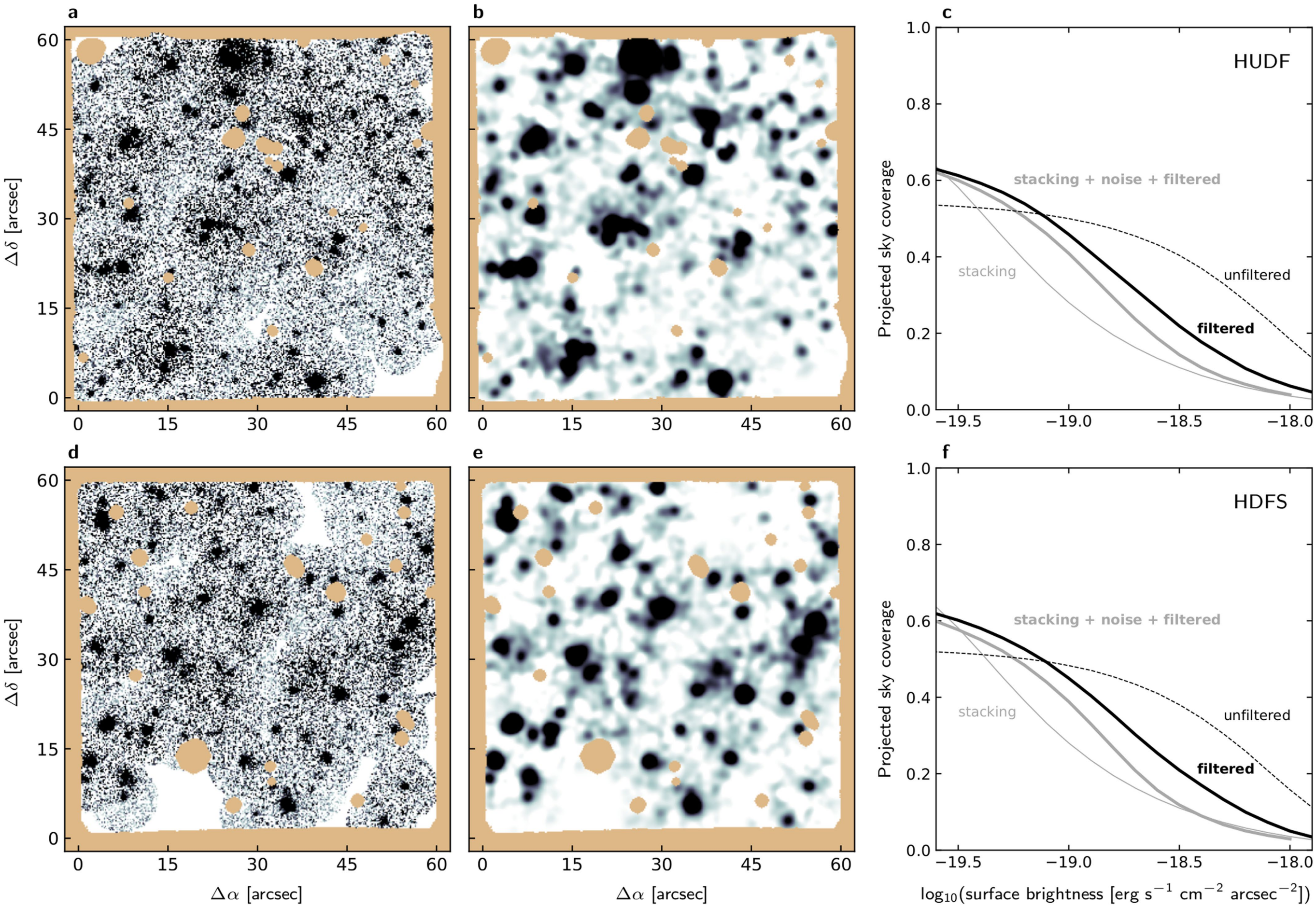
Data availability

The observations of the HUDF discussed in this paper were made using European Southern Observatory (ESO) Telescopes at the La Silla Paranal Observatory under programme IDs 094.A-0289, 095.A-0010, 096.A-0045 and 096.A-0045. The corresponding data are available on the ESO archive at http://archive.eso.org/cms.html. The data of the HDFS were obtained during MUSE commissioning observations and are available at http://muse-vlt.eu/science/hdfs-v1-0/.

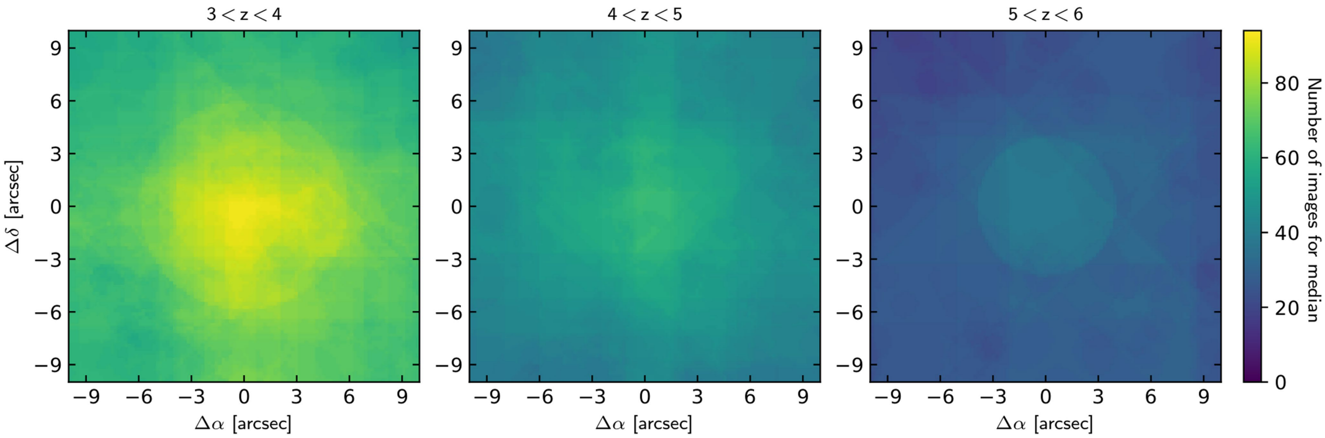
31. Bacon, R. et al. The MUSE Hubble Ultra Deep Field Survey. I. Survey description, data reduction, and source detection. *Astron. Astrophys*. **608**, A1 (2017).
32. Moffat, A. F. J. A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry. *Astron. Astrophys*. **3**, 455–461 (1969).
33. Herenz, E. C. & Wisotzki, L. LSDCat: Detection and cataloguing of emission-line sources in integral-field spectroscopy datacubes. *Astron. Astrophys*. **602**, A111 (2017).
34. Inami, H. et al. The MUSE Hubble Ultra Deep Field Survey. II. Spectroscopic redshifts and comparisons to color selections of high-redshift galaxies. *Astron. Astrophys*. **608**, A2 (2017).
35. Peng, C. Y., Ho, L. C., Impey, C. D. & Rix, H.-W. Detailed structural decomposition of galaxy images. *Astron. J*. **124**, 266–293 (2002).
36. Sersic, J. L. *Atlas de Galaxias Australes* Vol. 1 (Observatorio Astronomico, Cordoba, 1968).
37. Rahmati, A. et al. The distribution of neutral hydrogen around high-redshift galaxies and quasars in the EAGLE simulation. *Mon. Not. R. Astron. Soc*. **452**, 2034–2056 (2015).
38. Schaye, J. et al. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Mon. Not. R. Astron. Soc*. **446**, 521–554 (2015).
39. Rudie, G. C. et al. The gaseous environment of high-*z* galaxies: precision measurements of neutral hydrogen in the circumgalactic medium of $z \sim 2$–3 galaxies in the Keck Baryonic Structure Survey. *Astrophys. J*. **750**, 67 (2012).
40. Garel, T. et al. The UV, Lyman α, and dark matter halo properties of high-redshift galaxies. *Mon. Not. R. Astron. Soc*. **450**, 1279–1294 (2015).
41. Stark, D. P., Ellis, R. S., Chiu, K., Ouchi, M. & Bunker, A. Keck spectroscopy of faint $3 < z < 7$ Lyman break galaxies – I. New constraints on cosmic reionization from the luminosity and redshift-dependent fraction of Lyman α emission. *Mon. Not. R. Astron. Soc*. **408**, 1628–1648 (2010).
42. Caruana, J. et al. The MUSE-Wide survey: a measurement of the Ly α emitting fraction among $z > 3$ galaxies. *Mon. Not. R. Astron. Soc*. **473**, 30–37 (2018).
43. Stark, D. P., Ellis, R. S. & Ouchi, M. Keck spectroscopy of faint $3 > z > 7$ Lyman break galaxies: a high fraction of line emitters at redshift six. *Astrophys. J*. **728**, L2 (2011).



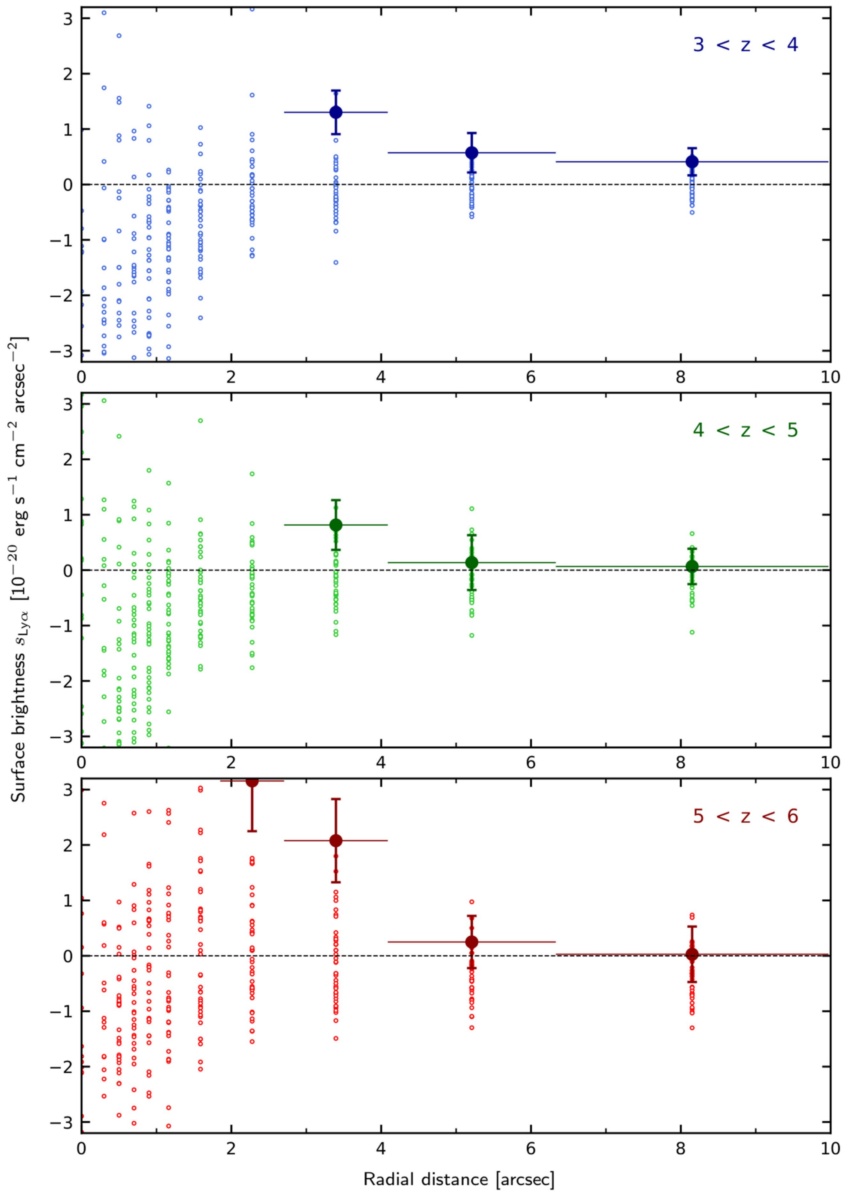
Extended Data Fig. 1 | Spatial distribution and redshifts of the Lyα emitter sample. a, The region observed with MUSE in the Hubble Ultra Deep Field (HUDF), **b**, the same for the Hubble Deep Field South (HDFS). Each Lyα emitter is represented by a circle colour-coded by redshift (key at right) and with a radius scaled by the integrated Lyα flux of the object. α, right ascension; δ, declination. The objects shown here constitute the full sample. There are several cases of significant crowding of unequal-redshift objects separated by less than a few arcseconds in projection. The underlying greyscale images show the two fields as seen with the HST.



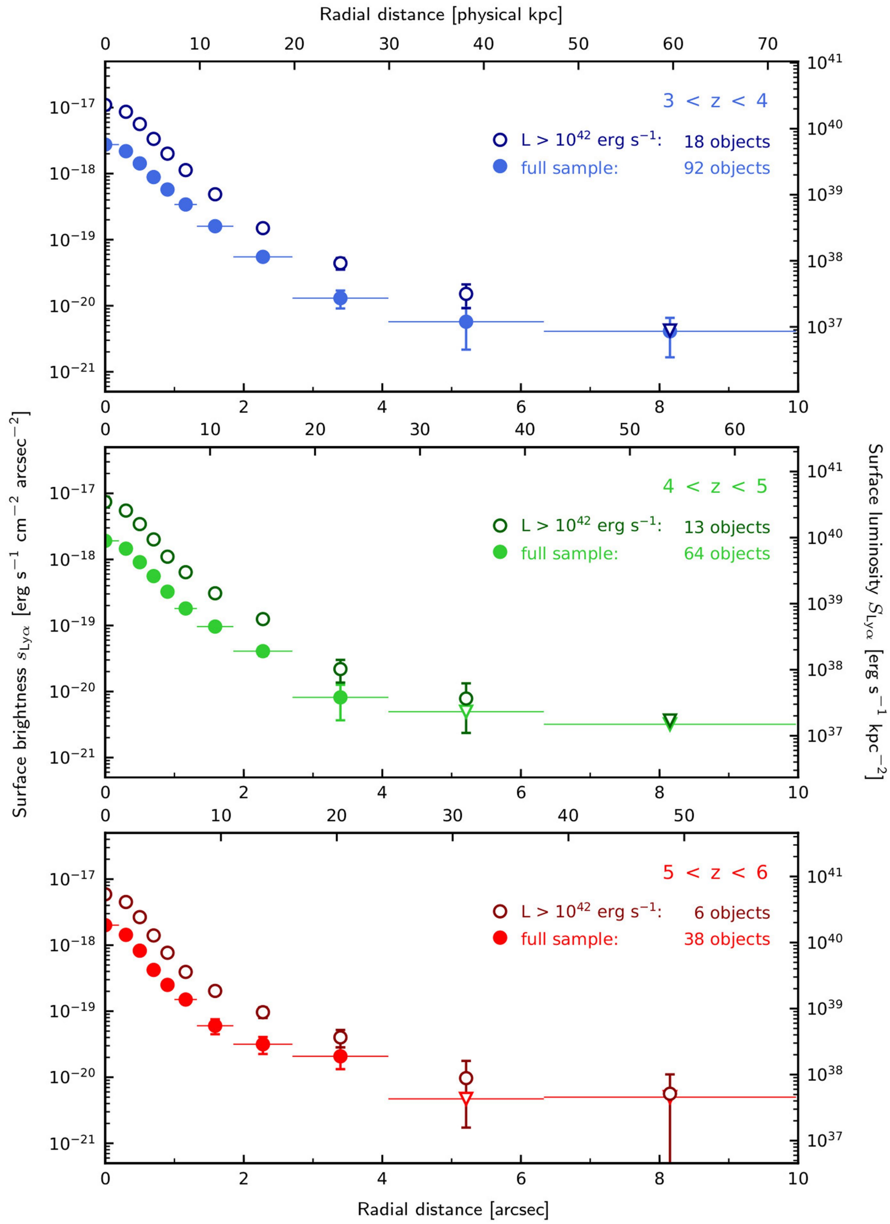
Extended Data Fig. 2 | Lyα sky coverage from direct projection. **a**–**c**, Data from HUDF; **d**–**f**, data from HDFS. Greyscale images display the projected and coadded Lyα emission over the redshift range $3 < z < 6$ separately for the two observed fields: **a**–**d**, without any spatial filtering; **b**–**e**, after Gaussian filtering with FWHM = 1.4″. The image **b** is identical to the blue overlay in Fig. 1. The light brown areas delineate the MUSE field of view and indicate masked bright foreground objects. **c**–**f**, the resulting fractional sky coverage; the black dashed line representing the unfiltered images and the thick black solid line representing the spatially filtered images. The thin grey line shows the result from the stacking analysis (Fig. 3) for comparison, and the thick grey line is a fiducial reconstruction of the sky coverage derived from a noisy and filtered version of the stacking-based model images. See the discussion in Methods for an interpretation of these figures.



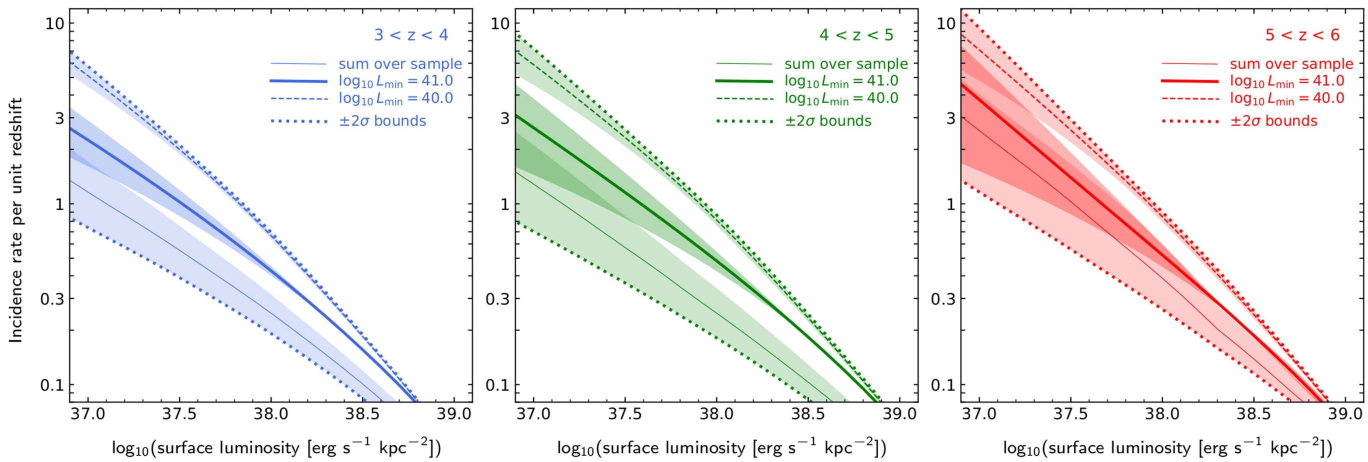
Extended Data Fig. 3 | Number of images contributing to the median stacks. **a**–**c**, Data for redshift ranges $3 < z < 4$ (**a**), $4 < z < 5$ (**b**) and $5 < z < 6$ (**c**). The colour code represents, for each pixel in a median-stacked image, the number of original image pixels contributing to it. This number differs from the total number of objects in a given stack because of the masks applied to several of the contributing images.



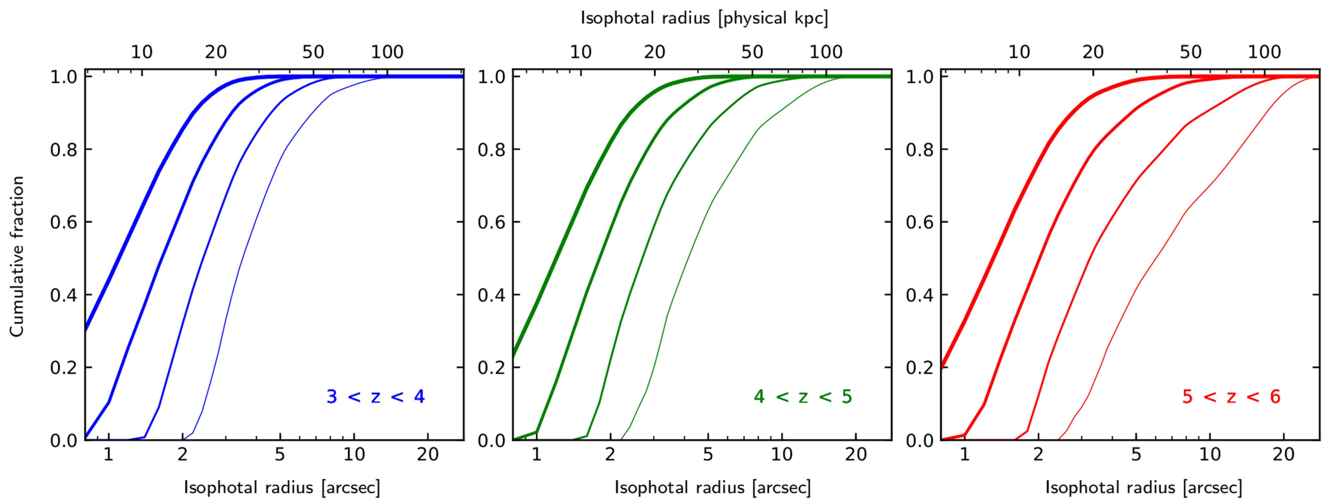
Extended Data Fig. 4 | Error calibration and sensitivity of azimuthally averaged profiles. Shown are 'zoom views' into the profiles (surface brightness versus radial distance) at very low surface brightnesses, with linear ordinate scaling so that negative measurements can be displayed. Data for three redshift ranges are shown: top, $3 < z < 4$; middle, $4 < z < 5$; and bottom, $5 < z < 6$. The small open circles represent the surface brightnesses measured in 40 realizations of a median-stacking analysis of empty regions as described in Methods. The filled symbols reproduce the data points from Fig. 2c. The horizontal bars again indicate the widths of the annuli, and the vertical error bars are based on the 1σ scatter of the empty field median-stack profiles; these were also adopted as error bars for the data points in Fig. 2c.



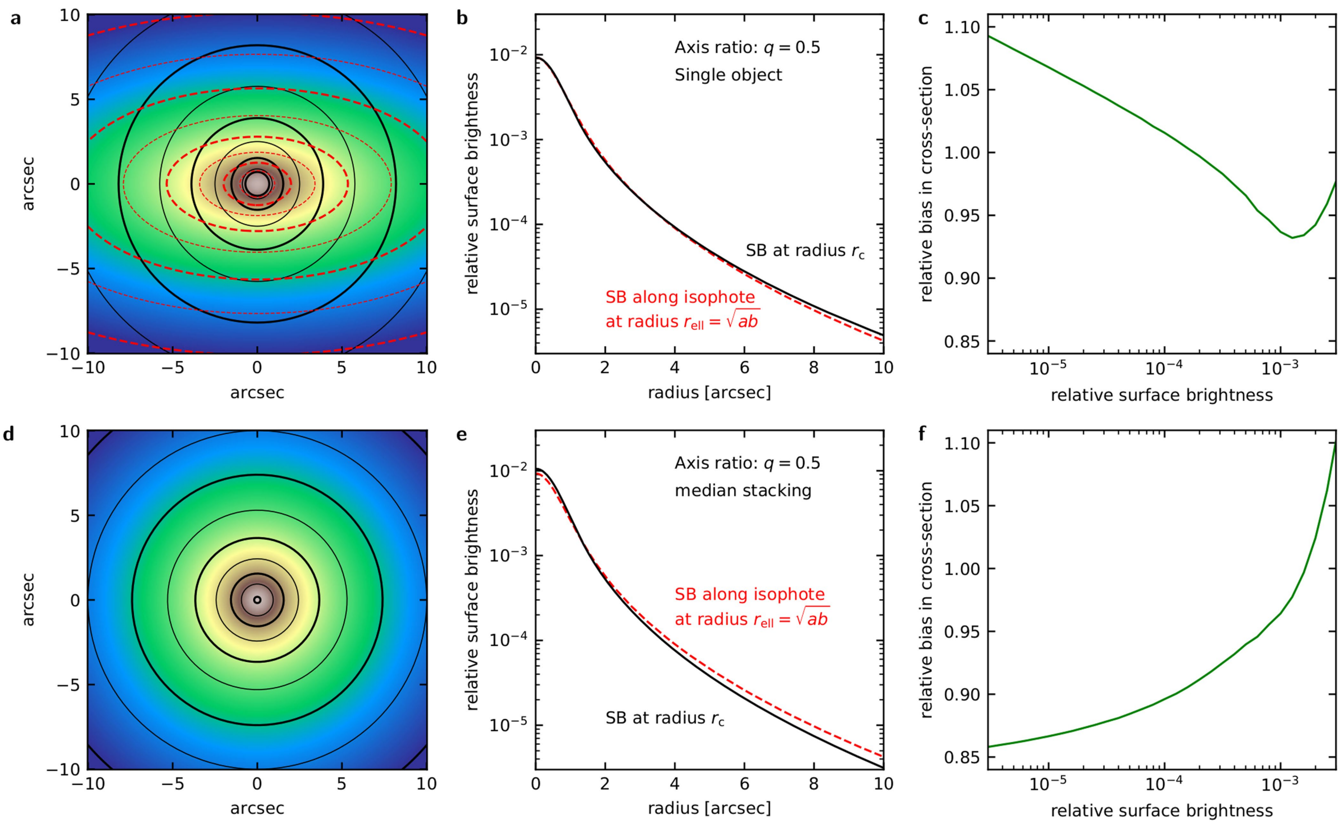
Extended Data Fig. 5 | A test of the self-similarity assumption for different Lyα luminosities. a–c, Comparison of azimuthally averaged radial profiles of median-stacked Lyα images above a minimum Lyα luminosity L (open circles) and with no such cut ('full sample', filled circles), for three redshift ranges (top, $3 < z < 4$; middle, $4 < z < 5$; and bottom, $5 < z < 6$). As in Fig. 2, the vertical bars on the data points quantify the 1σ surface brightness measurement errors, while the horizontal bars (drawn only for the filled symbols) indicate the widths of the annuli. Inverted triangles indicate upper limits. The right-hand ordinate provides the conversion from apparent surface brightnesses to redshift-corrected surface luminosities, evaluated at the central redshift of each bin.



Extended Data Fig. 6 | Comparison of approaches to determine the Lyα emission incidence rates. Each panel shows the cumulative incidence rate as a function of limiting surface luminosity for the specified redshift range (left, $3 < z < 4$; middle, $4 < z < 5$; and right, $5 < z < 6$), estimated by different methods: direct summation of Lyα cross-sections over the sample without correcting for incompleteness (equation (1) in the Methods section, thin lines), and integrating over the completeness-corrected luminosity function following equation (2) in the Methods section, using lower integration limits of $\ell_{min} = 41.0$ (best guess, thick solid line) and $\ell_{min} = 40.0$ (asymptotic case, dashed line), respectively. The shaded error bands for direct summation are dominated by field-to-field variance between the HUDF and the HDFS, with the upper envelope tracing the HUDF and the lower envelope tracing the HDFS results. For the luminosity function integration the error bands on these curves incorporate only the statistical uncertainties of the median-stacked profiles. The two thick dotted lines indicate the finally adopted lower and upper 2σ bounds on the 'best guess' results shown in Fig. 4.



Extended Data Fig. 7 | Cumulative contributions to incidence rates for different radii. Each panel (left, $3 < z < 4$; middle, $4 < z < 5$; and right, $5 < z < 6$) shows the cumulative fractional contributions of objects with different isophotal radii to the integrated Lyα emission cross-section dn/dz, using equation (2) with $\ell_{\rm min} = 41$. The four lines in each panel represent, from left to right and with decreasing line width, surface luminosity limits $\log_{10}[S_{\rm Ly\alpha}\ ({\rm erg\ s^{-1}\ kpc^{-2}})] = 38.5, 38, 37.5, 37$.



Extended Data Fig. 8 | Bias of estimated cross-sections if the emission is non-axisymmetric. a, Model image of an elongated surface brightness distribution, normalized to an integrated flux of 1, following an elliptical Sersic law with axis ratio $q = 0.5$ and smoothed with a Gaussian of 0.8″ FWHM. The colour code represents relative surface brightness (SB), and the red-dashed contours trace the isophotes at 0.5 dex separation. The black circles represent the radii where an azimuthally averaged profile over circular annuli gives the same surface SB values as the corresponding isophote. **b**, Radial profiles of the model image. The red-dashed line represents the input SB law as a function of generalized radius $r_{ell} = \sqrt{ab}$ where a, b are the major and minor axes of an isophote. The black line shows the profile obtained from azimuthal averaging over circular annuli against radius r_c. **c**, Ratios between the true isophotal cross-sections πab and those estimated from the circularized profile as πr_c^2 (that is, the ratios of the areas of the black circles and the corresponding red-dashed ellipses in **b**), as a function of surface brightness. **d**, Median-stacked image of an ensemble of 180 model objects with properties each as in **a**, but rotated in position angle between 0° and 180° in steps of 1°. The colour code again represents relative SB, and the black circles show the resulting isophotes at 0.5 dex separation. **e**, The black line traces the radial profile of the median-stacked image in **d**. The red-dashed line is the true elliptical SB distribution of a single object (same as in **b**). **f**, Ratios of cross-sections obtained from the median stack to the true isophotal ones in a single image, as a function of surface brightness.

Extended Data Table 1 | Values of the best-fit parameters for the analytic profiles

z range	F_{h}	$r_{\mathrm{eff,h}}$	n_{h}	F_{ps}	$\mathrm{FWHM}_{\mathrm{PSF}}$
3–4	1488 ± 83	0.86 ± 0.11	2.8 ± 1.1	232 ± 50	0.703
4–5	931 ± 82	0.90 ± 0.18	3.3 ± 1.9	150 ± 40	0.654
5–6	1002 ± 164	1.67 ± 0.86	6.5 ± 5.1	167 ± 34	0.606

These values were obtained by applying GALFIT to the median-stacked images. For each of the three adopted redshift ranges (column 1), the first three parameters characterize the circular Sersic model used to describe the Lyα haloes: halo flux F_{h} (in 10^{-20} erg s^{-1} cm^{-2}), effective radius $r_{\mathrm{eff,h}}$ (in arcsec), and Sersic index n_{h} (dimensionless), followed by the flux of the point-like component F_{ps} (same units as F_{h}). The quoted errors are 1σ uncertainty estimates. The last column provides the seeing (FWHM of the mean PSF in arcsec) at the appropriate wavelengths.

https://doi.org/10.1038/s41586-018-0524-1

An evolving jet from a strongly magnetized accreting X-ray pulsar

J. van den Eijnden[1]*, N. Degenaar[1], T. D. Russell[1], R. Wijnands[1], J. C. A. Miller-Jones[2], G. R. Sivakoff[3] & J. V. Hernández Santisteban[1]

Relativistic jets are observed throughout the Universe and strongly affect their surrounding environments on a range of physical scales, from Galactic binary systems[1] to galaxies and clusters of galaxies[2]. All types of accreting black hole and neutron star have been observed to launch jets[3], with the exception of neutron stars with strong magnetic fields[4,5] (higher than 10^{12} gauss), leading to the conclusion that their magnetic field strength inhibits jet formation[6]. However, radio emission recently detected from two such objects could have a jet origin, among other possible explanations[7,8], indicating that this long-standing idea might need to be reconsidered. But definitive observational evidence of such jets is still lacking. Here we report observations of an evolving jet launched by a strongly magnetized neutron star accreting above the theoretical maximum rate given by the Eddington limit. The radio luminosity of the jet is two orders of magnitude fainter than those seen in other neutron stars with similar X-ray luminosities[9], implying an important role for the properties of the neutron star in regulating jet power. Our result also shows that the strong magnetic fields of ultra-luminous X-ray pulsars do not prevent such sources from launching jets.

On 3 October 2017, the Neil Gehrels Swift Observatory (Swift) detected an outburst of a new X-ray transient Swift J0243.6+6124 (hereafter Sw J0243)[10]. The discovery of 9.86-s pulsations[11] identified this transient as an accreting pulsar: a relatively slowly spinning neutron star with a strong magnetic field ($B \geq 10^{12}$ G)[12], probably accreting from a high-mass companion Be star. Throughout its outburst, we observed this source at radio wavelengths over eight epochs with the Karl G. Jansky Very Large Array (VLA); see Fig. 1. After an initial non-detection early in the outburst, we detected significant (18.3σ) radio emission at 6 GHz close to the X-ray peak (Fig. 2), when the neutron star was accreting above the theoretical Eddington limit. The radio luminosity of the system subsequently decayed with the X-ray flux, while the radio spectral index α (where the flux density is $S_\nu \propto \nu^\alpha$ and ν is the frequency) gradually evolved throughout the outburst. We did not detect linearly polarized emission during any epoch, with a very stringent upper limit of about 15% during the third observation (see Extended Data Tables 1 and 2 for all measurements).

Its radio properties show that Sw J0243 launches an evolving jet. Whenever accreting compact objects launch steady jets, the radio and X-ray luminosity are coupled[9,13] (see Fig. 3), indicating a direct relationship between the X-ray-emitting accretion flow and the radio-emitting jet. After the initial radio non-detection, we observed such a coupling between the X-ray and radio luminosities of Sw J0243, with the radio luminosity decreasing as the X-ray luminosity of the outburst decayed. By estimating the correlation index between the 0.5–10-keV X-ray and 6-GHz radio luminosities, we measured $L_\mathrm{r} \propto L_\mathrm{X}^{0.54\pm0.16}$, which is consistent with both black-hole and neutron-star X-ray binaries[14] (see Methods).

The radio spectral shape and evolution also support a jet origin of the outburst. In radiofrequencies, jets launched from stellar-mass accretors emit synchrotron radiation with a spectral index that can vary over time, as observed in Sw J0243. The radio spectral index distribution of Sw J0243 starts out steep ($\alpha < 0$) and gradually evolves to a flat spectrum ($\alpha \geq 0$), as observed in canonical steady X-ray binary jets[15]. This systematic evolution during the outburst decay can be interpreted as follows: during the super-Eddington phase, where strong outflows are expected theoretically[16], discrete transient ejecta were launched. When the accretion rate decayed during the remainder of the outburst, the radio–X-ray correlation and the transition towards an inverted spectrum signalled that the radio emission arose from a compact, steady jet instead[15]. Alternatively, a gradual shift of the break frequency, where the jet spectrum transitions from optically thin to thick synchrotron radiation, could also be responsible for the observed evolution of the radio spectral index. As discussed in Methods, alternative physical or emission mechanisms cannot explain the observed combination of spectral index evolution, flux levels, radio–X-ray coupling and polarization. We note that both the observed polarization properties and the spectral shape and evolution rule out coherent radio pulsations being responsible for the radio emission.

Before our radio monitoring campaign of Sw J0243, jets had been confirmed in all types of X-ray binary system[4,9] except in strongly magnetized accreting pulsars, which are the most common X-ray binary type. Multiple large surveys in the 1970s and 1980s failed to detect radio emission from these sources[5,17,18], leading to the observational notion that their strong magnetic field prevents the formation of jets. Until recently, searches for radio emission from individual neutron stars with such field strengths also yielded non-detections[6], further strengthening this idea. As a result, strongly magnetized accreting neutron stars are often disregarded in theoretical studies of neutron-star jet formation[19].

Jet formation models developed for accreting neutron stars commonly invoke a magneto-centrifugal launch mechanism[6,20,21], in which the jet is launched by field lines anchored in the innermost accretion disk. Such models offer a straightforward theoretical explanation for the prevention of jet formation by strong magnetic fields: the neutron-star magnetosphere stops the formation of the inner accretion flow by dominating over the disk pressure[6], therefore preventing the launching of a jet. The first observational results to question this view were the recent radio detections[7,8] of the two strongly magnetized pulsars Her X-1 and GX 1+4. However, in contrast to our Sw J0243 observation, both sources were detected at a single frequency during a single epoch, meaning that the origin of the emission remained ambiguous. Given the lack of information on spectral shape, temporal evolution or coupling with the X-ray flux, a jet could neither be excluded nor directly inferred. Moreover, the properties of any putative jets—if present—could not be determined from the limited information available.

Our clear discovery of an evolving jet in Sw J0243 disproves the long-standing idea that strong magnetic fields prevent the launch of a jet. This directly indicates that existing models of jet formation in neutron-star X-ray binaries[6,20] need to be revisited. For instance, the jet-launching region must be much farther from the neutron star than in other classes of jet-forming systems. The presence of X-ray

[1]Anton Pannekoek Institute for Astronomy, University of Amsterdam, Amsterdam, The Netherlands. [2]International Centre for Radio Astronomy Research–Curtin University, Perth, Western Australia, Australia. [3]Department of Physics, University of Alberta, Edmonton, Alberta, Canada. *e-mail: a.j.vandeneijnden@uva.nl

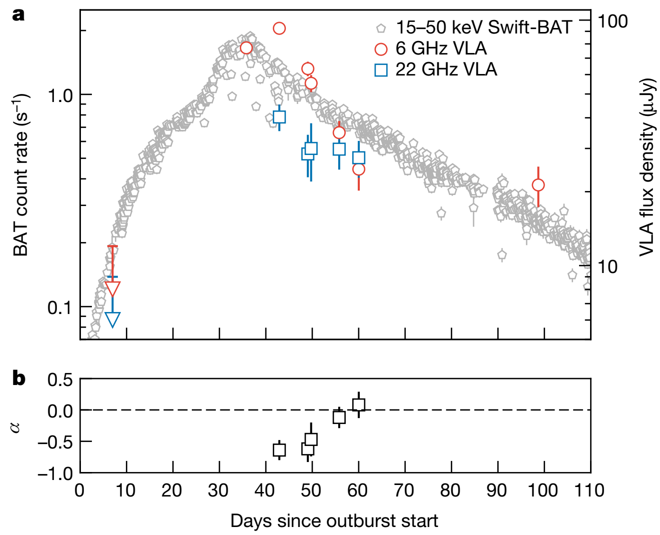


Fig. 1 | Radio and X-ray outburst light curve of Sw J0243. **a**, Radio flux densities detected by the VLA at 6 GHz and 22 GHz (red circles and blue squares; right axis) and the count rate between 15 keV and 50 keV measured by the Burst Alert Telescope (BAT) onboard Swift throughout the outburst (grey pentagons; left axis). Sw J0243 was not detected during the first radio epoch, marked by downward arrows. **b**, Radio spectral index α (flux density, $S_\nu \propto \nu^\alpha$) as a function of time. In both panels error bars are given at the 1σ level and upper limits are 3σ.

pulsations[11] shows that the magnetosphere dominates the inner accretion flow, channelling the material to the neutron-star poles. Conservatively estimated, its minimum size—at the outburst peak, during the first jet detection—corresponds to a magnetospheric radius of 320 gravitational radii (see Methods). Hence, the geometrically thin accretion disk must be truncated much farther away from the compact object than typically seen in X-ray binaries with weak magnetic fields ($\lesssim 10^9$ G), where the observed jets are thought to be launched close to the accretor[6,20,22]. Moreover, in strongly magnetized pulsars accreting at super-Eddington rates, such as Sw J0243, the magnetosphere might be completely enveloped by accreting material[23]. Such a configuration involves entirely different (geometric) properties of the inner accretion flow from those of other types of X-ray binary. However, the apparent coupling between the X-ray and radio luminosity during the decay and the spectral index evolution of Sw J0243 are similar to those of other black-hole and neutron-star X-ray binaries, but at much higher mass-accretion rates. Therefore, it is unclear what similarities exist in the jet formation mechanism and what role the magnetosphere has.

Despite the phenomenological similarities with jets from stellar-mass black holes and weakly magnetized neutron stars, the jet in Sw J0243 is orders of magnitude fainter in radio luminosity. This difference is evident in the L_X–L_r diagram shown in Fig. 3, where Sw J0243 falls two orders of magnitude below other neutron stars accreting at similar super-Eddington X-ray luminosities. Importantly, the only difference between Sw J0243 and these other neutron stars is that the latter have a weak magnetic field ($\lesssim 10^9$ G) and are spinning faster. Therefore, the difference in radio luminosity might suggest an important role for these fundamental properties of the neutron star in regulating jet power. This role fits with recent theoretical work[21] discussing a neutron-star jet model in which the jet is powered by the accretor's rotation, as in the Blandford–Znajek-type models for black holes[24], and not launched by field lines in the inner accretion disk, as in the magneto-centrifugal (Blandford–Payne-type) jet models commonly used for neutron stars[6,20].

This model, which was subsequently shown by numerical simulations[25] to be applicable to the super-Eddington accreting regime of Sw J0243, also predicts a suppression of two orders of magnitude in jet power for slowly pulsating, strongly magnetized accreting pulsars compared to their weakly magnetized, rapidly spinning counterparts.

Our discovery of a jet in a strongly magnetized accreting pulsar has two additional major implications. First, it implies that accreting pulsars form a large, hidden class of radio emitters, which are now accessible through the current generation of observatories with upgraded sensitivities. This unexplored population opens up new avenues to test general predictions of jet theory for all accreting systems. In Blandford–Znajek-type models of jet formation, a correlation would be expected between the spin and jet power[21,24]. This straightforward prediction has been difficult to test—estimates of the spin of black holes are challenging, and although pulsations provide an undisputed measure of neutron-star spins, the only neutron stars previously known to launch jets (those with weak magnetic fields) span merely a small range in spin frequency (a factor of about 5–6)[26]. By contrast, accreting pulsars with strong magnetic fields can span over three orders of magnitude

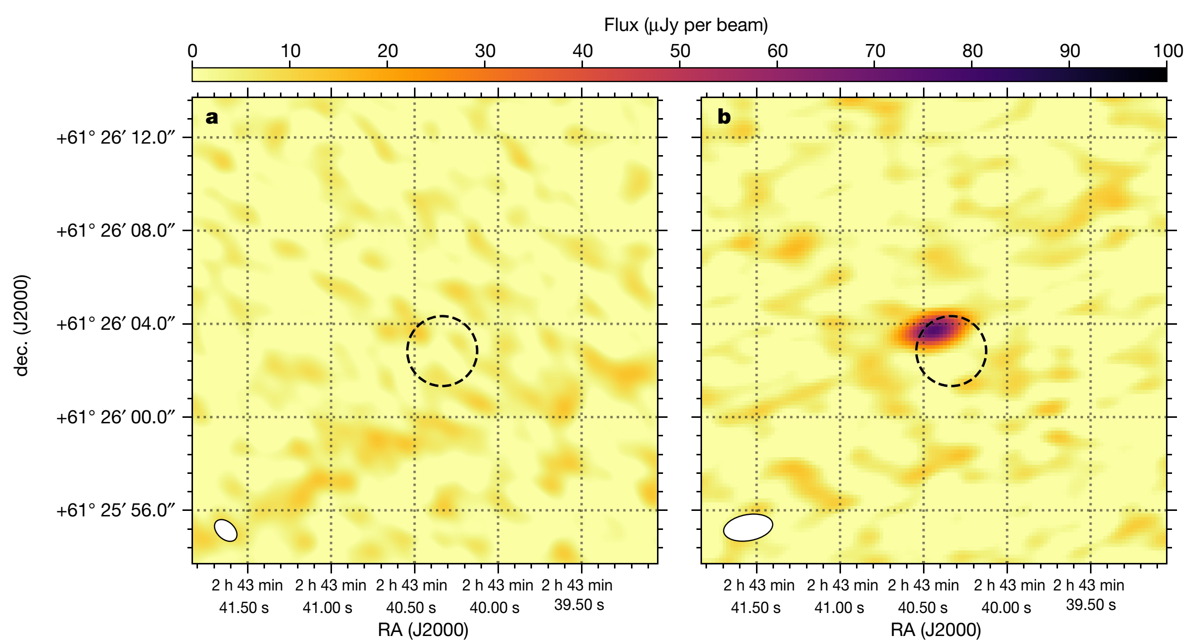


Fig. 2 | VLA detection images of Sw J0243. **a**, The 6-GHz image of the first VLA observation of Sw J0243. No 3σ-significant radio emission is observed within the black dashed circle, which indicates the 90% position contour of the X-ray Telescope (XRT) onboard Swift. **b**, The 6-GHz image of Sw J0243 during the second VLA epoch, when the target was first detected. A new, 18.3σ-significance source is coincident with the Swift-XRT position. The synthesized beam is shown in the bottom left corner of both panels.

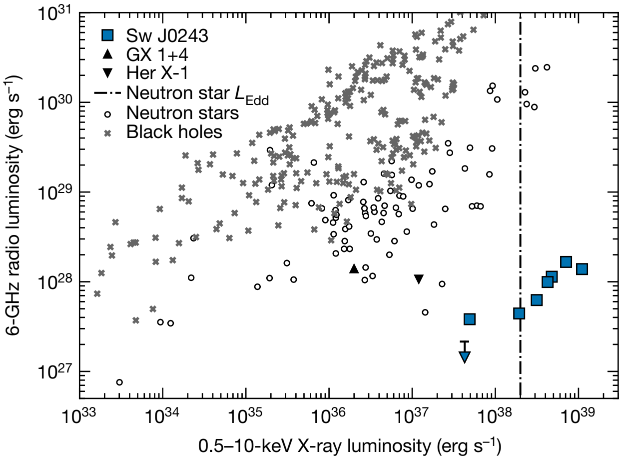


Fig. 3 | Radio and X-ray luminosities for X-ray binaries. The X-ray and radio luminosities of Sw J0243 during the eight epochs are shown, together with a large sample of accreting stellar-mass black holes and neutron stars. The dashed line shows the Eddington X-ray luminosity, L_{Edd}, for a $1.4M_{\odot}$ neutron star ($M_{\odot}$, mass of the Sun). See Methods for details on the sample shown and the estimation of the distance used. For visual clarity we do not plot any non-detections or uncertainties for any source in the comparison sample.

in spin and have similar and well measured magnetic fields. Now that we have found that strongly magnetized accreting pulsars can launch jets, future observational campaigns of this source class will probe the predicted relation between spin and jet power.

In addition, the detection of a jet in Sw J0243 expands the possible types of outflow in ultra-luminous X-ray sources (ULXs), which are binary systems with X-ray luminosities greatly exceeding the Eddington luminosity of a stellar-mass accretor. Super-Eddington winds have previously been observed in both black-hole and neutron-star ULXs[16], and jets have been inferred in a handful of black-hole ULXs through direct detection and the presence of surrounding bubbles[27]. Although several ULXs have been confirmed to be neutron stars through the detection of pulsations, recent theoretical[28] and observational[29] studies indicate that such strongly magnetized ULX pulsars could make up a large fraction of the population. Interestingly, the known ULX pulsars show similar X-ray behaviour to Galactic pulsars accreting from Be stars at super-Eddington rates[30], such as Sw J0243. Our detection of a jet in Sw J0243 therefore implies that, in addition to winds, ULX pulsars might also launch jets, unhampered by their strong magnetic fields.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0524-1.

Received: 14 February 2018; Accepted: 26 July 2018;
Published online 26 September 2018.

1. Gallo, E. et al. A dark jet dominates the power output of the stellar black hole Cygnus X-1. *Nature* **436**, 819–821 (2005).
2. Fabian, A. C. Observational evidence of active galactic nuclei feedback. *Annu. Rev. Astron. Astrophys.* **50**, 455–489 (2012).
3. Fender, R. Disc-jet-wind coupling in black hole binaries, and other stories. *Astron. Nachr.* **337**, 381–384 (2016).
4. Fender, R. P. & Hendry, M. A. The radio luminosity of persistent X-ray binaries. *Mon. Not. R. Astron. Soc.* **317**, 1–8 (2000).
5. Migliari, S., Ghisellini, G., Miller-Jones, J. & Russell, D. Jet models for neutron star X-ray binaries. *Int. J. Mod. Phys.* **8**, 108–113 (2012).
6. Massi, M. & Kaufman Bernadó, M. Magnetic field upper limits for jet formation. *Astron. Astrophys.* **477**, 1–7 (2008).
7. van den Eijnden, J. et al. Radio emission from the X-ray pulsar Her X-1: a jet launched by a strong magnetic field neutron star? *Mon. Not. R. Astron. Soc.* **473**, L141–L145 (2018).
8. van den Eijnden, J. et al. Discovery of radio emission from the symbiotic X-ray binary system GX 1+4. *Mon. Not. R. Astron. Soc.* **474**, L91–L95 (2018).
9. Migliari, S. & Fender, R. P. Jets in neutron star X-ray binaries: a comparison with black holes. *Mon. Not. R. Astron. Soc.* **366**, 79–91 (2006).
10. Kennea, J. A., Lien, A. Y., Krimm, H. A., Cenko, S. B., & Siegel, M. H. Swift J0243.6+6124: swift discovery of an accreting NS transient. *Astron. Telegr.* 10809 (2017).
11. Jenke, P. & Wilson-Hodge, C. A. Fermi GBM detects pulsations from Swift J0243.6+6124. *Astron. Telegr.* 10812 (2017).
12. Doroshenko, V., Tsygankov, S. & Santangelo, A. Orbit and intrinsic spin-up of the newly discovered transient X-ray pulsar Swift J0243.6+6124. *Astron. Astrophys.* **613**, A19 (2018).
13. Merloni, A., Heinz, S. & di Matteo, T. A fundamental plane of black hole activity. *Mon. Not. R. Astron. Soc.* **345**, 1057–1076 (2003).
14. Gallo, E., Degenaar, N. & van den Eijnden, J. Hard state neutron star and black hole X-ray binaries in the radio: X-ray luminosity plane. *Mon. Not. R. Astron. Soc.* **478**, L132–L136 (2018).
15. Fender, R. P., Belloni, T. M. & Gallo, E. Towards a unified model for black hole X-ray binary jets. *Mon. Not. R. Astron. Soc.* **355**, 1105–1118 (2004).
16. Kosec, P. et al. Evidence for a variable ultrafast outflow in the newly discovered ultraluminous pulsar NGC 300 ULX-1. *Mon. Not. R. Astron. Soc.* **479**, 3978–3986 (2018).
17. Duldig, M. L. et al. Radio identifications of 11 X-ray sources at 2 CM. *Mon. Not. R. Astron. Soc.* **187**, 567–580 (1979).
18. Nelson, R. F. & Spencer, R. E. A search for radio emission from X-ray binaries and related objects. *Mon. Not. R. Astron. Soc.* **234**, 1105–1118 (1988).
19. Kato, Y., Mineshige, S. & Shibata, K. Magnetohydrodynamic accretion flows: formation of magnetic tower jet and subsequent quasi-steady state. *Astrophys. J.* **605**, 307–320 (2004).
20. Blandford, R. D. & Payne, D. G. Hydromagnetic flows from accretion discs and the production of radio jets. *Mon. Not. R. Astron. Soc.* **199**, 883–903 (1982).
21. Parfrey, K., Spitkovsky, A. & Beloborodov, A. M. Torque enhancement, spin equilibrium, and jet power from disk-induced opening of pulsar magnetic fields. *Astrophys. J.* **822**, 33 (2016).
22. Gandhi, P. et al. An elevation of 0.1 light-seconds for the optical jet base in an accreting Galactic black hole system. *Nat. Astron.* **1**, 859–864 (2017).
23. Mushtukov, A. A., Suleimanov, V. F., Tsygankov, S. S. & Ingram, A. Optically thick envelopes around ULXs powered by accreting neutron stars. *Mon. Not. R. Astron. Soc.* **467**, 1202–1208 (2017).
24. Blandford, R. D. & Znajek, R. L. Electromagnetic extraction of energy from Kerr black holes. *Mon. Not. R. Astron. Soc.* **179**, 433–456 (1977).
25. Parfrey, K. & Tchekhovskoy, A. General-relativistic simulations of four states of accretion onto millisecond pulsars. *Astrophys. J.* **851**, L34 (2017).
26. Patruno, A., Haskell, B. & Andersson, N. The spin distribution of fast-spinning neutron stars in low-mass X-Ray binaries: evidence for two subpopulations. *Astrophys. J.* **850**, 106 (2017).
27. Kaaret, P., Feng, H. & Roberts, T. P. Ultraluminous X-ray sources. *Annu. Rev. Astron. Astrophys.* **55**, 303–341 (2017).
28. King, A. & Lasota, J.-P. ULXs: neutron stars versus black holes. *Mon. Not. R. Astron. Soc.* **458**, L10–L13 (2016).
29. Walton, D. J. et al. Evidence for pulsar-like emission components in the broadband ULX sample. *Astrophys. J.* **856**, 128 (2018).
30. Tsygankov, S. S., Doroshenko, V., Lutovinov, A. A., Mushtukov, A. A. & Poutanen, J. SMC X-3: the closest ultraluminous X-ray source powered by a neutron star with non-dipole magnetic field. *Astron. Astrophys.* **605**, A39 (2017).

Acknowledgements We thank the VLA for rapidly accepting our proposal and performing the Director's Discretionary Time (DDT) radio observations, and E. Gallo for providing the black hole sample used in Fig. 3. J.v.d.E., N.D. and J.V.H.S. appreciate support from a Netherlands Organisation for Scientific Research (NWO) Vidi grant awarded to N.D. T.D.R. is supported by an NWO Veni grant. R.W. is supported by an NWO Top grant. J.C.A.M.-J. is supported by an Australian Research Council Future Fellowship (FT140101082). G.R.S. acknowledges support from an NSERC Discovery grant. This work used data supplied by the UK Swift Science Data Centre at the University of Leicester. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. This work used data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by the national institutions participating in the Gaia Multilateral Agreement.

Reviewer information *Nature* thanks R. Fender and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.v.d.E. led the two VLA DDT observing campaigns, performed the analysis of the Swift data and wrote the manuscript, with comments from all authors. J.v.d.E., N.D. and T.D.R. designed the radio monitoring strategy. T.D.R. and J.v.d.E. jointly analysed the VLA radio data. J.V.H.S. calculated the Gaia DR2 distance estimate. All authors made contributions to the scientific case and commented on multiple versions of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0524-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.v.d.E.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Radio observations. We observed Sw J0243 with the VLA over eight epochs between 10 October 2017 and 9 January 2018. The observations were part of two Director's Discretionary Time programmes, VLA/17B-406 and VLA/17B-420, for the first two and remaining six epochs, respectively. The total observing time was 13 h. In two observations, we observed the target only at the C band, centred at 6 GHz with 4 GHz of bandwidth. In the other six observations, we observed both at the C band (with the same setup) and K band, with the latter centred at 22 GHz with 8 GHz of bandwidth. Detailed information about each epoch can be found in Extended Data Table 1.

In all epochs, the primary calibrator was J0137+331 (3C48) and the nearby phase calibrator was J0244+6228 (1.04° angular separation from the target). When included in the setup (epochs 3–7), the leakage calibrator was J0319+4130 (3C84). For the target field, the centre was offset by 6″ in the north direction from the detection position[10] of the XRT onboard Swift to prevent possible correlator artefacts at the phase centre from affecting the results. During all observations, the VLA was in its B configuration. More detailed information, such as beam sizes and position angles in each observing band and epoch, is given in Extended Data Table 1.

To analyse the observations, we used the Common Astronomy Software Application package[31] (CASA) v4.7.2 to flag, calibrate and image the data. We removed radio-frequency interference using a combination of automated flagging routines and careful visual inspection. Given the lack of bright radio emission in the target field, we did not self-calibrate. Using the multi-frequency multi-scale CLEAN task with Briggs weighting and a robustness of 1 (to reduce the effects of the side-lobes of a neighbouring source), we imaged Stokes I at all observed frequencies for all epochs, and Stokes Q and U at 6 GHz for epochs with leakage calibration. We did not image Stokes Q and U at 22 GHz because we did not detect any linearly polarized emission at 6 GHz. Therefore, no such emission is expected at 22 GHz, and the better r.m.s. sensitivity at 6 GHz yields tighter upper limits.

Accreting X-ray binaries are expected to be unresolved point sources for the VLA. Therefore, we determined fluxes by fitting an elliptical Gaussian equalling the beam size to the source in the image plane. We measured the RMS of the cleaned image over a region close to the target position. We determined a single flux density in each band and, owing to the faintness of the radio emission, did not divide the C- and K-band frequency ranges further. A quick check for time variability did not reveal any evidence for substantial variability within observations.

The target was not detected in our first observational epoch, with 3σ upper-limits on the flux densities of 12 μJy per beam and 9 μJy per beam in the C and K band, respectively. Sw J0243 was detected in all following observations. All flux densities are listed in Extended Data Table 2. The radio position of Sw J0243, measured at 6 GHz from the first detection, is RA = 02 h 43 min 40.440 s ± 0.029 s and dec. = +61° 26′ 03.73″ ± 0.10″.

All positions determined from the radio detections are consistent with the Swift-XRT X-ray position in every epoch. In Extended Data Fig. 1, we show the target field during the initial non-detection and the first detection. The combination of the spatial coincidence between the X-ray and radio position and the coupled X-ray and radio variability shows that the observed radio emission originates from Sw J0243.

In epochs with both C- and K-band observations, we calculated the spectral index to investigate the spectral shape. The power-law spectral index α (where the flux density is $S_\nu \propto \nu^\alpha$) between two frequencies ν_1 and ν_2 with corresponding flux densities S_1 and S_2 is calculated as:

$$\alpha = \frac{\log(S_1/S_2)}{\log(\nu_1/\nu_2)}$$

To calculate the uncertainty on the spectral index for each individual epoch, we propagate the uncertainties on the measured flux densities and the range in frequencies through a Monte Carlo simulation: in each iteration, two new flux densities are drawn from Gaussian distributions centred on the measured flux densities with standard deviations equalling the measured uncertainties. New frequencies are drawn from uniform distributions over the frequency range of each band. The resulting calculated spectral index is then saved. After 10^6 iterations, we calculate the spectral index uncertainty as the standard deviation of the simulated spectral indices.

X-ray flux measurements. For the study of Sw J0243 in the X-ray luminosity–radio luminosity plane, accurate and precise X-ray fluxes during the radio epochs are required. Three X-ray instruments consistently observed the entire outburst of Sw J0243: the XRT and Burst Alert Telescope (BAT) instruments onboard Swift[32] and the Monitor of All-sky X-ray Image[33] (MAXI) onboard the International Space Station. The BAT and MAXI are monitoring instruments, whereas XRT exposures are pointed observations. Both monitoring instruments only provide count rates of observed targets, which cannot be converted to a flux straightforwardly without knowing the shape of the X-ray spectrum. The comparison of XRT fluxes and monitoring count rates shows that the broadband X-ray spectral shape evolves during the outburst. This implies that the count-rate-to-flux conversion for the BAT and MAXI is also variable and therefore makes both instruments inconvenient for accurately estimating the X-ray flux. The MAXI is additionally unsuitable because a visual inspection of the light curve shows several unphysical jumps in count rate, pointing towards systematic errors in the monitoring.

The above considerations make the XRT the most reliable instrument to determine the X-ray flux of Sw J0243 during the radio epochs. Five out of the eight radio epochs had quasi-simultaneous XRT coverage (within 2 days). For the remaining three epochs, such XRT observations were not available. However, preliminary flux estimates for all XRT observations, extracted using the Swift-XRT data products generator[34] (http://www.swift.ac.uk/user_objects/), show that Sw J0243 decayed in a steady, log–linear fashion as a function of time. Therefore, for the three radio epochs without close XRT coverage, we estimated the logarithm of the X-ray flux using linear interpolation between the logarithmic fluxes of the preceding and subsequent XRT pointings. Before describing the actual X-ray flux measurements, we stress that the BAT count rate of Sw J0243 between 15 keV and 50 keV also decayed in a log–linear fashion during our radio monitoring. This implies that the XRT observations, which only provide spectra up to 10 keV, are representative of both the soft- and hard-X-ray decay of Sw J0243.

We extracted spectra from the radio position of Sw J0243 using the Swift-XRT data products generator[34] and used XSPEC[35] v12.9.0u to fit the data and determine the fluxes. All analysed observations were taken in the window-timing mode. We did not use the fluxes provided by the data products generator for our actual measurements; these fluxes are based on a power-law-only model, which is not necessarily accurate for every spectrum. Moreover, the automatic fits are performed between 0.3 keV and 10 keV, whereas the window-timing mode of XRT is subject to calibration uncertainties for moderately-to-heavily absorbed sources, possibly resulting in poor fits at low energies (see, for example, http://www.swift.ac.uk/analysis/xrt/digest_cal.php#abs).

We fitted each spectrum with a model containing interstellar absorption, a blackbody component and a power law (TBABS*(BBODYRAD+PO)). Because Be/X-ray binaries can have strongly variable local absorption, we did not tie the absorption column between spectra. We assumed Wilms abundances[36] and Verner cross-sections[37] and fitted the spectra in the reliable energy range (0.7–10 keV). We then determined unabsorbed fluxes and their uncertainties in the 0.5–10 keV range using CFLUX and the best-fitting model. Information on the analysed observations and the fluxes determined in this analysis, including interpolated fluxes, are listed in Extended Data Table 3. The best-fit parameters for each spectrum are listed in Extended Data Table 4.

Gaia distance measurement. We used the recent Gaia Data Release 2[38,39] to obtain an independent measurement of the distance to the system. The measured parallax of Sw J0204 is $\pi = 0.0952 \pm 0.0302$ mas. We followed the standard Bayesian method to infer the distance towards the system[40]. The likelihood function assumes a normal distribution for the Gaia parallaxes and a suggested prior distribution modelled as an exponential decreasing volume density function, with a length scale of 1.35 kpc corresponding to the line-of-sight value[41]. We took into account the zero point from the global astrometric solution $\alpha_{zp} = -0.029$ mas[42] and we used a Markov chain Monte Carlo procedure (as implemented in *emcee*[43]) to sample the posterior distribution of the distance. The marginal posterior distributions are shown in Extended Data Fig. 1. We found a median value of $D = 7.3$ kpc with 16th and 84th percentiles of 6.1 kpc and 8.9 kpc, respectively. We stress that the posterior distribution is not symmetric and caution should therefore be exercised in using these numbers.

Given the large fractional error of the parallax, the shape of the posterior distribution deviates from a Gaussian distribution and the upper tail is very sensitive to the choice of the prior distribution. We investigated the robustness of our distance estimate with different choices of prior distributions, as shown in Extended Data Fig. 1. When using a uniform prior with a maximum distance of 50 kpc, the median of the distribution shifts towards larger distances. However, the lower limit of the distance is greater than 5.0 kpc at ≥99% confidence level for both priors. Therefore, the Gaia measurement shows that the source is located at a distance of at least 5.0 kpc, independent of the prior used. We conservatively adopt this lower limit on the distance in Fig. 3.

During the peak of the outburst, around the time of the first radio detection (epoch 2), the XRT unabsorbed flux at 0.5–10 keV is $(3.69 \pm 0.03) \times 10^{-7}$ erg s^{-1} cm^{-2}. For a conservative (prior-independent) minimum distance to the source of 5 kpc, this flux corresponds to an X-ray luminosity of $1.1 \times 10^{39} \times [D/(5\ \text{kpc})]^2$ erg s^{-1}. If we apply a bolometric correction, by extrapolating the best-fitting model to the 0.1–100 keV range, we find an even higher luminosity of $1.5 \times 10^{39} \times [D/(5\ \text{kpc})]^2$ erg s^{-1}. The theoretical Eddington luminosity of an accreting neutron star is 2×10^{38} erg s^{-1}. This shows that even for its closest estimated distance, Sw J0243 firmly reached the super-Eddington regime during the outburst.

Swift-BAT light curve. To show the long-term X-ray evolution of Sw J0243, we display the Swift-BAT light curve in Fig. 1; however, for clarity, we show a cleaned

version of this light curve. Owing to the extremely high count rates of the source, the measured count rate sometimes dropped by an order of magnitude in individual exposures—the BAT team ascribes these drops to software issues and not to intrinsic variability in Sw J0243 (https://swift.gsfc.nasa.gov/results/transients/weak/SwiftJ0243.6p6124/). Therefore, we masked these anomalously low points in the light curve, which occur between 19 and 60 days into the outburst, around the times with the highest count rates. In this time interval the actual rates exceeded 0.7 counts per second, so we removed all exposure with lower count rates. We stress that this cleaning procedure is for visual purposes only and does not affect our actual measurements or conclusions.

Estimating the magnetospheric radius. The magnetospheric radius is defined as the radius where the pressure of the magnetosphere and accreting material are equal. Therefore, this radius will depend on the strength of the magnetic field, B, and the rate of accretion. The latter can be estimated from the bolometric flux, F, the distance, D (together providing the X-ray luminosity), the accretion efficiency, η (which converts the mass-accretion rate to luminosity) and an anisotropy correction factor, f, which accounts for the anisotropy of the emitted X-rays. Finally, the type of accretion (that is, wind or disk) has to be taken into account through a geometrical correction factor, k. For standard neutron-star parameters—a mass of $1.4M_{\odot}$ and a radius of 10 km—the magnetospheric radius $R_{\rm m}$ (in gravitational radii, $R_{\rm g}=GM/c^2$, where G is the gravitational constant, M is the mass and c is the speed of light in vacuum) can be estimated from the above parameters as[44–46]:

$$R_{\rm m}=k\left(\frac{B}{1.2\times10^5\,{\rm G}}\right)^{4/7}\left(\frac{f}{\eta}\right)^{-4/14}\left(\frac{F}{10^{-9}\,{\rm erg\,s^{-1}\,cm^{-2}}}\right)^{-4/14}\left(\frac{D}{5\,{\rm kpc}}\right)^{-4/7}R_{\rm g}$$

Although not all parameters are known precisely, we can use this equation to estimate a minimum size of the magnetosphere during the outburst. The maximum unabsorbed, bolometric X-ray flux observed by Swift-XRT around a radio epoch, which will give the smallest magnetospheric radius, is 4.9×10^{-7} erg s^{-1} cm^{-2} (but see below). Gaia measurements with an exponential prior distribution imply a median distance estimate of 7.3 kpc, which we adopt for this calculation—this provides a more conservative lower limit on $R_{\rm m}$ than using a minimum distance of 5 kpc. The minimum value of k is 0.5, as appropriate for disk accretion[47]. The accretion efficiency is typically assumed to be 0.1, and the anisotropy correction is close to unity[45]. Finally, the magnetic field is not measured directly but can be determined from the X-ray pulsations[12] to exceed 10^{12} G. Combining these numbers yields $R_{\rm m}\gtrsim320R_{\rm g}$ (that is, about 670 km).

Following typical assumptions, we used the bolometric X-ray flux, combined with an efficiency of 10%, to probe the mass-accretion rate that balances the magnetospheric pressure. However, a non-negligible fraction of the X-ray flux might be emitted from the neutron-star surface with higher efficiency, which would imply that this approach might overestimate the mass-accretion rate. On the other hand, outflows from the neutron star or disc would cause the flux-derived mass-accretion rate to be underestimated. Given these contradictory possibilities, we did not correct for either of these processes: correcting for the former would lead to a larger magnetospheric radius, which is already consistent with our approach of calculating a lower limit; correcting for the latter would lead to a lower radius—but this correction is small, given the weak scaling between mass-accretion rate and magnetospheric radius ($-2/7$). Thus, correcting for either case does not affect our conclusions.

Radio–X-ray correlation sample. For the radio–X-ray correlation, shown in Fig. 3, we use a comprehensive sample of hard-state Atoll neutron-star sources and hard-state black holes from the large body of observational studies of X-ray binaries performed over the past decades. This sample is freely available online (https://jakobvdeijnden.wordpress.com/radioxray/) and was originally compiled for a different study focusing on the radio–X-ray luminosity plane of accreting neutron stars[14]. To this sample, we added Z sources[9], two jet-quenched accreting neutron stars[48,49] and the accreting pulsars[7,8] GX 1+4 and Her X-1. We added these sources as interesting comparisons with Sw J0243: the Z sources have similar X-ray luminosities, the jet-quenched neutron stars have similar radio luminosities and the accreting pulsars have similar physical characteristics. We note that, as discussed extensively in the next section, it remains unclear whether the radio emission from these two accreting pulsars originates from a jet.

The radio luminosities in the full sample were collected at 5 GHz, whereas we measured the 6-GHz radio luminosity of Sw J0243. Hence, we transformed the 5-GHz sample luminosities to 6-GHz ones by assuming a flat spectrum, which amounts to multiplying all luminosities in the sample by 6/5. The assumption of a flat radio spectrum is not accurate for all observations. For instance, a clear effect of the radio spectral shape on the position of black-hole systems on the radio–X-ray luminosity plane has recently been demonstrated[50]. However, making this simplifying assumption is valid because we use the large sample only for a broad qualitative comparison between Sw J0243 and other types of source. Our conclusions—that Sw J0243 shows an apparent coupling between in- and outflow and is two orders of magnitude fainter than the Z sources—are not affected by assuming a flat radio spectrum.

Finally, we note that we plot the 0.5–10 keV X-ray luminosity of Sw J0243 to be consistent with the full sample. Before Sw J0243, no radio emission confirmed to be from a jet had been detected from any confirmed high-mass X-ray binary system containing a neutron star. Therefore, all neutron stars in the sample reside in low-mass X-ray binaries. While the 0.5–10-keV X-ray luminosity does not necessarily probe the same components of the accretion flow in low- and high-mass X-ray binaries, we plot this energy range to remain consistent between all sources and with the existing literature.

Measuring the X-ray–radio correlation index. We measured the correlation index from the 0.5–10-keV X-ray and 6-GHz radio luminosities in epochs 2–8 (that is, those with radio detections). We fit the following function to these seven data points:

$$L_{\rm r}=L_{\rm r,ref}\left(\frac{L_{\rm X}}{L_{\rm X,ref}}\right)^{\beta}$$

where $L_{\rm X,ref}$ is the average X-ray flux of all epochs and $L_{\rm r,ref}$ and β are free parameters. We find $\beta=0.54\pm0.16$, which is consistent with the indices for both the black-hole and weakly magnetized neutron-star X-ray binaries[14].

It is important to treat this value with caution. Our monitoring result for Sw J0243 spans a factor of approximately 20 in X-ray luminosity and 5 in radio luminosity during the outburst. However, to accurately measure the coupling index between the radio and X-ray luminosities, detailed monitoring over at least two orders of magnitude in X-ray luminosity is strongly recommended[51]. Therefore, although our result is consistent with other X-ray binaries, the exact value is not necessarily representative of the entire outburst or accreting pulsars in general.

From the $L_{\rm X}$–$L_{\rm r}$ diagram, it is clear that without including the radio detection with the lowest X-ray luminosity, the correlation index distribution would be steeper. Although we cannot draw conclusions based on a single data point, this might reflect changes in the jet properties as the source becomes sub-Eddington and the accretion flow geometry changes.

Alternative interpretations. Here we briefly discuss a few alternative interpretations for the observed radio properties of Sw J0243. As mentioned in the main text, none of these alternative explanations can account for the observed combination of radio–X-ray coupling, flux levels, spectral index evolution and polarization properties.

First, the stellar wind in a high-mass X-ray binary system can emit in radio frequencies. Through a combination of optically thick and thin free–free processes, the radio spectrum of such a wind could be flat[52] (that is, $\alpha=0$), as we observe in later epochs (see Fig. 1 and Extended Data Table 1). However, the systematic evolution seen in the spectral index, which is similar to that in low-mass X-ray binaries, is not expected for a stellar wind. The same goes for the clear coupling between radio and X-ray flux.

We can also consider the flux levels expected from a stellar wind. The typical flux S_{ν} of a stellar wind can be estimated[7,8,52,53] for a given mass-accretion rate $\dot{M}$, velocity v, distance D and observing frequency ν:

$$S_{\nu}=7.26\left(\frac{\nu}{10\,{\rm GHz}}\right)^{0.6}\left(\frac{\dot{M}}{10^{-6}M_{\odot}\,{\rm yr}^{-1}}\right)^{4/3}\left(\frac{v}{100\,{\rm km\,s^{-1}}}\right)^{-4/3}\left(\frac{D}{1\,{\rm kpc}}\right)^{-2}\,{\rm mJy}$$

where we ignore the electron temperature owing to its negligible effect on the predicted flux and assume a hydrogen wind (which yields the highest predicted flux). Conservatively assuming the escape velocity of a typical Be star as a minimum for the wind velocity, and using the lower limit of 5 kpc on the distance (which yields the highest flux density) at a frequency of 6 GHz, we find that the mass-loss rate in the wind needs to exceed $10^{-5}M_{\odot}$ yr^{-1} to account for the observed flux levels around the outburst peak. Such rates are only associated with Wolf–Rayet stars and are highly unlikely for a Be star, which are more likely to lose mass at a maximum rate of[54] $10^{-9}M_{\odot}$ yr^{-1}. At rates of $10^{-9}M_{\odot}$ yr^{-1}, the wind flux would not be expected to exceed 0.01 μJy—orders of magnitude below our radio detections.

Alternatively, accreting neutron stars could launch an outflow through the propeller mechanism. If the rotational velocity of the accreting material is lower than the neutron-star spin at the magnetospheric radius, the material can be expelled in a propeller outflow[55]. However, given the magnetic field strength and spin of Sw J0243, such an outflow is not expected at the high mass-accretion rates that are present when we detect radio emission[56], as the magnetospheric radius is then pushed far inside the co-rotation radius. Instead, the propeller regime and its associated outflows are typically expected to be below 10^{36} erg s^{-1} for Be/X-ray binary systems[57], which is over two orders of magnitude below the super-Eddington X-ray luminosities of Sw J0243.

Radio pulsations at the neutron-star spin frequency also cannot be the origin of the observed emission. Although Sw J0243 is too faint to explicitly search for pulsations at the known spin, the spectral shape and evolution rule out this origin: radio pulsations have a steep ($\alpha \approx -1.4$) spectrum that does not evolve[58], in contrast to the different, evolving spectral shape observed in Sw J0243.

Coherent emission of any form is ruled out owing to the lack of observed circular polarization in Sw J0243 in any epoch.

Finally, shocks between the accreting material and the magnetosphere could give rise to radio emission. However, although the luminosity resulting from this mechanism could be expected to scale with the accretion rate, and thus the X-ray luminosity, we do not necessarily expect the shock spectrum to evolve as observed: the regular evolution of the spectrum from optically thin to thick, coupled to the decaying X-rays, implies that the same mechanism is responsible for all emission. The spectral shape towards the end of our radio monitoring (that is, flat) is inconsistent with the optically thin spectrum expected for the shocked emission.

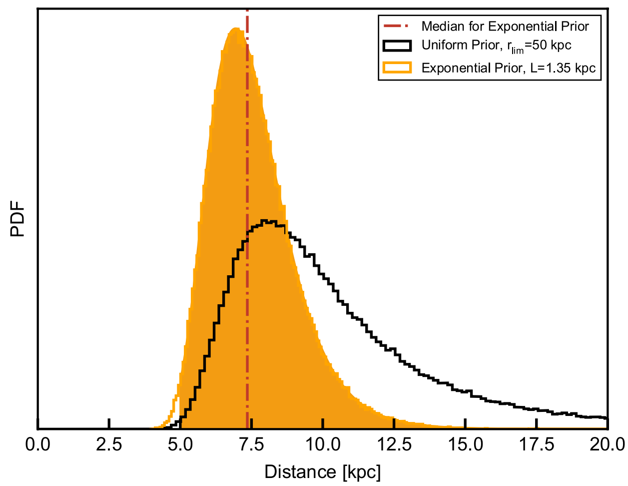
Therefore, none of these alternative mechanisms can account for our radio observations of Sw J0243. The observed radio properties directly point towards a jet origin (as argued in the main text). Combined, the exclusion of alternatives and direct implication of a jet origin make Sw J0243 a completely distinct case from Her X-1 and GX 1+4. Although those strongly magnetized accreting neutron stars were recently detected in radio frequencies, these single-frequency and single-epoch detections could not directly imply a jet origin[7,8]. Several of the alternative mechanisms discussed above could also not be excluded. Therefore, although inspiring for our multi-band monitoring campaign of Sw J0243, those detections could neither convincingly prove the presence of jets in strongly magnetized neutron stars (thus disproving the existing theory) nor provide details on the properties of such jets.

Code availability. The code used to estimate the distance from the Gaia DR2 measurements is available at https://github.com/Alymantara/Sw_J0243. All data analysis software is publicly available for download (CASA: https://casa.nrao.edu; HEASoft: https://heasarc.nasa.gov/lheasoft/). This research used Astropy, a community-developed core Python package for Astronomy[59], available at https://www.astropy.org.

Data availability

The VLA observations analysed in this work will become publicly available in the NRAO Science Data Archive (https://archive.nrao.edu/archive/advquery.jsp) on 8 November 2018 (first two epochs) and 20 February 2019 (remaining epochs), under project codes 17B-406 and 17B-420, respectively. However, prior access to the VLA observations will be granted by the corresponding author upon reasonable request. All Swift X-ray data are accessible in the HEASARC data archive. The radio–X-ray correlation data sample is available online at https://github.com/jvandeneijnden/XRB-Lx-Lr-Sample.

31. McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. CASA architecture and applications. *ASP Conf. Ser.* **376**, 127–130 (2007).
32. Gehrels, N. et al. The Swift gamma-ray burst mission. *Astrophys. J.* **611**, 1005–1020 (2004).
33. Matsuoka, M. et al. The MAXI mission on the ISS: science and instruments for monitoring all-sky X-ray images. *Publ. Astron. Soc. Jpn* **61**, 999–1010 (2009).
34. Evans, P. A. et al. Methods and results of an automatic analysis of a complete sample of Swift-XRT observations of GRBs. *Mon. Not. R. Astron. Soc.* **397**, 1177–1201 (2009).
35. Arnaud, K. A. XSPEC: the first ten years. *ASP Conf. Ser.* **101**, 17–20 (1996).
36. Wilms, J., Allen, A. & McCray, R. On the absorption of X-rays in the interstellar medium. *Astrophys. J.* **542**, 914–924 (2000).
37. Verner, D. A., Ferland, G. J., Korista, K. T. & Yakovlev, D. G. Atomic data for astrophysics. II. New analytic FITS for photoionization cross sections of atoms and ions. *Astrophys. J.* **465**, 487–498 (1996).
38. Gaia Collaboration. The Gaia mission. *Astron. Astrophys*. **595**, A1 (2016).
39. Gaia Collaboration. Gaia Data Release 2. Summary of the contents and survey properties. *Astron. Astrophys*. **616**, A1 (2018).
40. Bailer-Jones, C. A. L. Estimating distances from parallaxes. *Publ. Astron. Soc. Pacif.* **127**, 994 (2015).
41. Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G. & Andrae, R. Estimating distances from parallaxes IV: distances to 1.33 billion stars in Gaia Data Release 2. *Astronom. J.* **156**, 58 (2018).
42. Lindegren, L. et al. Gaia Data Release 2: the astrometric solution. *Astron. Astrophys*. **616**, A2 (2018).
43. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: the MCMC hammer. *Publ. Astron. Soc. Pacif.* **125**, 306–312 (2013).
44. Pringle, J. E. & Rees, M. J. Accretion disc models for compact X-ray sources. *Astron. Astrophys*. **21**, 1–9 (1972).
45. Ibragimov, A. & Poutanen, J. Accreting millisecond pulsar SAX J1808.4-3658 during its 2002 outburst: evidence for a receding disc. *Mon. Not. R. Astron. Soc*. **400**, 492–508 (2009).
46. Cackett, E. M. et al. Broad relativistic iron emission line observed in SAX J1808.4-3658. *Astrophys. J.* **694**, L21–L25 (2009).
47. Long, M., Romanova, M. M. & Lovelace, R. V. E. Locking of the rotation of disk-accreting magnetized stars. *Astrophys. J.* **634**, 1214–1222 (2005).
48. Miller-Jones, J. C. A. et al. Evolution of the radio–X-ray coupling throughout an entire outburst of Aquila X-1. *Astrophys. J.* **716**, L109–L114 (2010).
49. Gusinskaia, N. V. et al. Jet quenching in the neutron star low-mass X-ray binary 1RXS J180408.9-342058. *Mon. Not. R. Astron. Soc*. **470**, 1871–1880 (2017).
50. Espinasse, M. & Fender, R. P. Spectral differences between the jets in 'radio-loud' and 'radio-quiet' hard-state black hole binaries. *Mon. Not. R. Astron. Soc*. **473**, 4122–4129 (2018).
51. Corbel, S. et al. The 'universal' radio/X-ray flux correlation: the case study of the black hole GX 339-4. *Mon. Not. R. Astron. Soc*. **428**, 2500–2515 (2013).
52. Wright, A. E. & Barlow, M. J. The radio and infrared spectrum of early-type stars undergoing mass loss. *Mon. Not. R. Astron. Soc.* **170**, 41–51 (1975).
53. Körding, E. et al. A transient radio jet in an erupting dwarf nova. *Science* **320**, 1318–1320 (2008).
54. Snow, T. P. Jr. Stellar winds and mass-loss rates from Be stars. *Astrophys. J.* **251**, 139–151 (1981).
55. Illarionov, A. F. & Sunyaev, R. A. Why the number of galactic X-ray stars is so small? *Astron. Astrophys*. **39**, 185–195 (1975).
56. Campana, S. The quiescent X-ray emission of three transient X-ray pulsars. *Astrophys. J.* **580**, 389–393 (2002).
57. Tsygankov, S. S. et al. Propeller effect in two brightest transient X-ray pulsars: 4U 0115+63 and V 0332+53. *Astron. Astrophys*. **593**, A16 (2016).
58. Bates, S. D., Lorimer, D. R. & Verbiest, J. P. W. The pulsar spectral index distribution. *Mon. Not. R. Astron. Soc.* **431**, 1352–1358 (2013).
59. The Astropy Collaboration et al. The Astropy Project: building an inclusive, open-science project and status of the v2.0 core package. *Astron. J.* **156**, 123 (2018).



Extended Data Fig. 1 | Marginal posterior distributions for the distance to Sw J0243. We show the distribution for an exponential and a uniform prior. The median value (50th percentile) of the distribution for the exponential prior is shown as the dot-dashed line. L is the scale parameter of the exponential prior and r_{lim} is the maximum distance in the uniform prior. PDF, probability density function.

Extended Data Table 1 | Overview of VLA radio observations of Sw J0243

Radio Epoch	Start (UTC)	End (UTC)	Observing frequencies	Leakage Calibrator	Beam size (position angle)
1	2017-10-10 05:37:00	2017-10-10 06:09:50	6 GHz 22 GHz	No No	1.12” x 0.74” (47.4 deg) 0.30” x 0.20” (55.3 deg)
2	2017-11-08 00:46:52	2017-11-08 01:29:42	6 GHz	No	2.13” x 1.12” (-80.0 deg)
3	2017-11-15 03:21:21	2017-11-15 04:46:36	6 GHz 22 GHz	Yes No	1.32” x 0.99” (35.3 deg) 0.40” x 0.29” (52.2 deg)
4	2017-11-21 06:27:33	2017-11-21 07:52:46	6 GHz 22 GHz	Yes No	1.42” x 0.97” (-30.9 deg) 0.34” x 0.28” (-17.1 deg)
5	2017-11-22 23:29:24	2017-11-23 00:54:36	6 GHz 22 GHz	Yes No	1.89” x 1.06” (-87.6 deg) 0.57” x 0.33” (-74.1 deg)
6	2017-11-28 00:26:51	2017-11-28 01:58:04	6 GHz 22 GHz	Yes No	1.54” x 1.02” (70.7 deg) 0.45” x 0.27” (82.3 deg)
7	2017-12-02 05:47:30	2017-12-02 07:20:14	6 GHz 22 GHz	Yes No	1.33” x 1.03” (-31.2 deg) 0.35” x 0.28” (-12.8 deg)
8	2018-01-09 22:14:52	2018-01-09 22:57:44	6 GHz	No	1.61” x 1.10” (76.9 deg)

For each radio epoch, we list the start and end time of the target observations in UTC (that is, not including the initial setup and calibration), the observing frequencies, whether we observed a leakage calibrator, and the beam size and position angle (in degrees east of north) at each frequency. The 6-GHz observations were performed with 4 GHz of bandwidth and the 22-GHz observations with 8 GHz of bandwidth.

Extended Data Table 2 | VLA radio flux density, polarization and position measurements

Radio Epoch	Observing frequency	Flux density [μJy]	Spectral index α	Linear polarisation	6 GHz position
1	6 GHz 22 GHz	< 12.0 < 9.0	-	-	-
2	6 GHz	77.1 ± 4.2	-	-	RA: 02:43:40.440 ± 0.029s Dec: +61:26:03.73 ± 0.10"
3	6 GHz 22 GHz	92.6 ± 3.8 40.3 ± 5.0	-0.64 ± 0.16	< 17%	RA: 02:43:40.425 ± 0.022s Dec: +61:26:03.73 ± 0.18"
4	6 GHz 22 GHz	63.4 ± 4.3 28.5 ± 5.6	-0.62 ± 0.21	< 27%	RA: 02:43:40.419 ± 0.015s Dec: +61:26:03.80 ± 0.13"
5	6 GHz 22 GHz	55.3 ± 4.4 30.0 ± 8.0	-0.47 ± 0.27	< 34%	RA: 02:43:40.430 ± 0.026s Dec: +61:26:03.74 ± 0.10"
6	6 GHz 22 GHz	34.8 ± 4.0 29.8 ± 5.2	-0.12 ± 0.17	< 47%	RA: 02:43:440 ± 0.024s Dec: +61:26:03.65 ± 0.13"
7	6 GHz 22 GHz	24.7 ± 4.5 27.5 ± 4.7	0.08 ± 0.21	< 75%	RA: 02:43:40.419 ± 0.028s Dec: +61:26:03.69 ± 0.23"
8	6 GHz	21.3 ± 4.0	-	-	RA: 02:43:40.432 ± 0.042s Dec: +61:26:03.97 ± 0.21"

For each radio epoch and observing frequency, we show the observed flux densities (or 3σ upper limits in case of non-detection), the spectral index when both 6- and 22-GHz observations were carried out, the most stringent upper limit on linear polarization per epoch, if available, and the 6-GHz position per epoch. All uncertainties are 1σ, while upper limits are quoted at 3σ. The errors on the position are calculated by taking the maximum of the synthesized beam size divided by the signal-to-noise ratio of the source detection and 10% of the synthesized beam size, following VLA guidelines.

Extended Data Table 3 | Swift-XRT flux measurements

Radio Epoch	Swift XRT ObsId(s)	Start date(s)	Unabsorbed flux [10^{-8} erg s^{-1} cm^{-2}]	Interpolated flux [10^{-8} erg s^{-1} cm^{-2}]
1	10336007	2017-10-10	1.43 ± 0.01	n/a
2	10336022	2017-11-09	36.9 ± 0.25	n/a
3	10336025	2017-11-15	23.7 ± 0.16	n/a
4	10336025 10336031	2017-11-15 2017-11-27	23.7 ± 0.16 10.5 ± 0.07	16.0 ± 0.11
5	10336025 10336031	2017-11-15 2017-11-27	23.7 ± 0.16 10.5 ± 0.07	14.2 ± 0.10
6	10336031	2017-11-27	10.5 ± 0.07	n/a
7	10336033	2017-12-01	6.47 ± 0.04	n/a
8	10467007 10467008	2018-01-02 2018-01-13	2.04 ± 0.01 1.47 ± 0.01	1.64 ± 0.01

For each radio epoch, we list the Swift-XRT observations used to determine the unabsorbed X-ray flux. When two observations (ObsIds) are listed, the X-ray flux estimate for that radio epoch was determined through log–linear interpolation between the two observations. Three leading zeros have been removed from all ObsIds. All errors are quoted at 1σ.

Extended Data Table 4 | Swift-XRT spectral fit parameters

Swift XRT ObsId	N_H [10^{22} cm^{-2}]	T_{BB} [keV]	N_{BB}	Γ	N_{PO} [phot/keV/cm^2/s]	χ^2 / ν
10336007	1.60 ± 0.10	1.94 ± 0.06	43 ± 8	1.96 ± 0.16	1.85 ± 0.26	1028.8 / 865
10336022	1.19 ± 0.08	2.08 ± 0.06	1036 ± 162	1.84 ± 0.15	39.2 ± 4.6	814.3 / 877
10336025	1.28 ± 0.08	1.96 ± 0.05	810 ± 104	1.99 ± 0.14	28.7 ± 3.2	967.5 / 874
10336031	1.57 ± 0.09	1.96 ± 0.04	387 ± 46	2.13 ± 0.16	13.3 ± 1.7	944.6 / 873
10336033	1.46 ± 0.07	1.96 ± 0.04	212 ± 24	1.77 ± 0.13	6.56 ± 0.65	1007.4 / 889
10467007	1.47 ± 0.08	1.75 ± 0.06	80 ± 10	1.81 ± 0.13	2.42 ± 0.25	930.4 / 876
10467008	1.42 ± 0.04	-	-	1.31 ± 0.02	1.48 ± 0.04	1058.2 / 870

For each analysed Swift-XRT observation, we list the best-fit spectral parameters for a TBABS*(BBODYRAD+POWERLAW) model in XSPEC. N_H is the neutral hydrogen density, T_{BB} and N_{BB} are the blackbody temperature and normalization, respectively, and Γ and N_{PO} are the power-law index and normalization, correspondingly. The final column lists the χ^2 divided by the number of degrees of freedom, ν. As in Be/X-ray binaries, local absorption can contribute to the total absorption column and we do not tie N_H between observations. In observation 10467008, the inclusion of a blackbody spectral component was not statistically required. All errors are quoted at 1σ.

LETTER

https://doi.org/10.1038/s41586-018-0533-0

Glider soaring via reinforcement learning in the field

Gautam Reddy[1,5], Jerome Wong-Ng[1,5], Antonio Celani[2], Terrence J. Sejnowski[3,4] & Massimo Vergassola[1]*

Soaring birds often rely on ascending thermal plumes (thermals) in the atmosphere as they search for prey or migrate across large distances[1–4]. The landscape of convective currents is rugged and shifts on timescales of a few minutes as thermals constantly form, disintegrate or are transported away by the wind[5,6]. How soaring birds find and navigate thermals within this complex landscape is unknown. Reinforcement learning[7] provides an appropriate framework in which to identify an effective navigational strategy as a sequence of decisions made in response to environmental cues. Here we use reinforcement learning to train a glider in the field to navigate atmospheric thermals autonomously. We equipped a glider of two-metre wingspan with a flight controller that precisely controlled the bank angle and pitch, modulating these at intervals with the aim of gaining as much lift as possible. A navigational strategy was determined solely from the glider's pooled experiences, collected over several days in the field. The strategy relies on on-board methods to accurately estimate the local vertical wind accelerations and the roll-wise torques on the glider, which serve as navigational cues. We establish the validity of our learned flight policy through field experiments, numerical simulations and estimates of the noise in measurements caused by atmospheric turbulence. Our results highlight the role of vertical wind accelerations and roll-wise torques as effective mechanosensory cues for soaring birds and provide a navigational strategy that is directly applicable to the development of autonomous soaring vehicles.

In reinforcement learning, an animal maximizes its long-term reward by taking actions in response to its external environment and internal state. Learning occurs by reinforcing behaviour based on feedback from past experiences. Similar ideas have been used to develop intelligent agents that have achieved spectacular performance in strategic games such as backgammon[8] and Go[9], visual-based video game play[10] and robotics[11,12]. In the field, however, constraints imposed by variable and uncontrolled conditions prevent learning agents from using data-intensive learning algorithms and the optimization of model design needed for quicker learning. These are the conditions most often faced by living organisms.

A striking example in nature is provided by thermal soaring. Atmospheric convection is not consistent across days and, even under suitable conditions, the locations, sizes, durations and strengths of nearby thermals are unpredictable. As a result, the distribution of samples used to train the glider differs day-to-day. Gliders and birds operate at spatial and temporal scales where fluctuations in wind velocities are due to turbulent eddies lasting a few seconds that may mask or falsely enhance a glider's estimate of its mean climb rate. Further, the measurement of navigational cues using standard instrumentation may be consistently biased by aerodynamic effects that require precise quantification. Here, we demonstrate that reinforcement learning can meet the challenge of learning to soar effectively in atmospheric turbulent environments. In past work, by contrast, the manoeuvring of an autonomous helicopter in ref. [11] is a control problem that is decoupled from environmental fluctuations and has little trial-to-trial variability. Past autonomous soaring algorithms have largely relied on locating the centroid of a drifting Gaussian thermal[13–16], which is unrealistic, or have applied learning methods in highly simplified simulated settings[17–19].

Using the reinforcement learning framework[7], we may describe the behaviour of the glider as an agent traversing different states (s) by taking actions (a) while receiving a local reward (r). The goal is to find a behavioural policy that maximizes the 'value': that is, the mean sum of future rewards up to a specified horizon. We seek a model-free approach, which estimates the value of different actions at a particular state (called the Q function) solely through the agent's experiences during repeated instances of the task, thereby bypassing the modelling of complex atmospheric physics and aerodynamics (see Methods). The optimal policy is subsequently derived by taking actions with the highest Q value at each state, where the state includes sensorimotor cues and the glider's aerodynamic state.

To identify mechanosensory cues that could guide soaring, we recently combined the above ideas with simulations of virtual gliders in numerically generated turbulent flow[20]. Two cues emerged from our screening: (1) the vertical wind acceleration (a_z) along the glider's path and (2) the spatial gradients in the vertical wind velocity across the wings of the glider (ω). Intuitively, the two cues correspond to the gradient of the vertical wind velocity in the longitudinal and lateral directions of the glider, which locally orient it towards regions of higher lift. Simulations[20] further showed that the glider's bank angle is the crucial aerodynamic control variable; additional variables such as the angle of attack, or other mechanosensory cues such as temperature or vertical velocity, offer minor improvements when navigating within a thermal.

To learn to soar in the field, a glider (wingspan, 2 m) was equipped with autonomous soaring capabilities (Fig. 1a, b). The glider is equipped with a flight controller, which uses a feedback control system to modulate the glider's ailerons and elevator such that the bank angle and pitch take the values desired by the behavioural policy being used (we use two different behavioural policies during initial learning, and the gliders then implement a further policy—the final navigational strategy—after learning). Relevant measurements, such as the altitude, ground velocity ($\boldsymbol{u}$), airspeed, bank angle (μ) and pitch, are made continuously at 10 Hz with standard instrumentation (see Methods). At fixed time intervals, the glider changes its heading by modulating its bank angle in accordance with the implemented behavioural policy.

Noise and biases that affect learning in the field require the development of appropriate methods to extract environmental cues from measurements made by sensory devices. We found that estimates of a_z from the derivative of the vertical ground velocity (u_z) are biased by longitudinal motions of the glider about the pitch axis as the glider responds to an imbalance of forces and moments while turning. By modelling the glider's longitudinal dynamics, we obtain an unbiased estimate of the local vertical wind velocity (w_z), and a_z as its derivative (see Methods). The estimation of the spatial gradients across the wings, ω, poses a greater challenge, as it involves the difference between two noisy measurements at relatively close positions. The key observation that we used here is that the glider rolls because of contributions from vertical wind velocity gradients, the feedback control mechanism and various aerodynamic effects. The resulting roll-wise torque can be estimated from the small deviations of the true bank angle from

[1]Department of Physics, University of California, San Diego, La Jolla, CA, USA. [2]The Abdus Salam International Center for Theoretical Physics, Trieste, Italy. [3]The Salk Institute for Biological Studies, La Jolla, CA, USA. [4]Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. [5]These authors contributed equally: Gautam Reddy, Jerome Wong-Ng.
*e-mail: massimo@physics.ucsd.edu

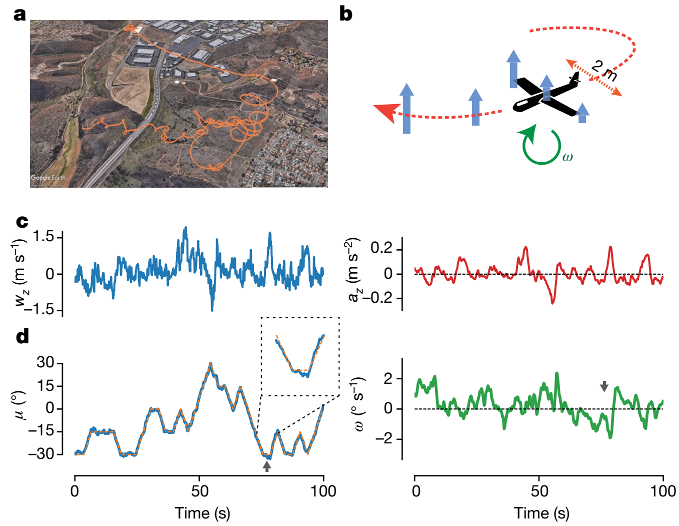


Fig. 1 | Soaring in the field by using turbulent navigational cues. **a**, A trajectory (orange line) of our glider soaring in Poway, California. **b**, A cartoon of the glider showing the available navigational cues—gradients in vertical wind velocities (indicated by the length of the blue arrows) along the trajectory and across its wings, which generate a vertical wind acceleration a_z and a roll-wise torque ω, respectively. **c**, A sample trace of the estimated vertical wind velocity w_z and corresponding a_z obtained in the field. **d**, The measured bank angle μ and the estimated ω during the same trial as in **c**. ω (solid green line) is estimated from the small deviations of the measured bank angle (solid blue line) from the expected bank angle (dashed orange line) after accounting for other effects (see Methods). The black arrows mark the enlarged bank angle trajectory shown in the inset in the left panel.

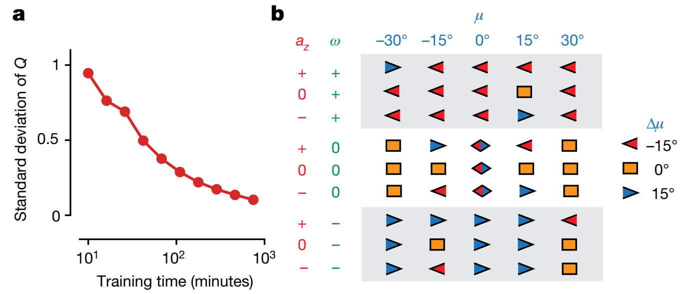


Fig. 2 | Convergence of the learning algorithm and the learned strategy for navigating thermal plumes. **a**, The convergence of Q values during learning as measured by the standard deviation of the mean Q value versus training time in the field, obtained by bootstrapping from the experiences accumulated up to that point. **b**, The final learned policy. Each symbol corresponds to the best action (increasing or decreasing the bank angle μ by 15° or maintaining the same μ, as shown in the key on the right) to be taken when the glider observes a particular (a_z, ω) pair and is banked at μ. Combined symbols depict pairs of actions that are equally rewarding. A positive (negative) ω corresponds to a higher vertical wind velocity on the left (right) wing of the glider and a positive (negative) μ corresponds to turning right (left) with respect to the glider's heading.

the desired one, and a new dynamical model allows us to separate the ω contribution due to velocity gradients from the other effects (see Methods). A sample trace of the resulting unbiased estimate of ω is shown in Fig. 1c, d, together with traces of w_z, μ and unbiased estimates of a_z.

Equipped with a proper procedure for estimating environmental cues, we next addressed the specifics of learning in the field. First, to constrain our state space, we discretized the range of values of a_z and ω into three states each: positive high (+), neutral (0) and negative high (−). Second, we found that learning is accelerated by choosing a_z attained at the subsequent time step as the reward signal. The choice of a_z (rather than w_z) is an instance of reward shaping that is justified in Supplementary Information, where we show that using a_z as a reward still leads to a policy that optimizes the long-term gain in height. This property is a special case of our general result that a particular reward function or its time derivatives (of any order) yield the same optimal policy (Supplementary Information). Choosing w_z as the reward fails to drive learning in the soaring problem, possibly because the velocities (and thus the rewards) are correlated across states and their temporal statistics strongly deviates from the Markovianity assumption in reinforcement learning methods[7]. Velocity fluctuations in turbulent flow are long-correlated: that is, their correlation timescale is determined by the largest timescale of the flow (see, for instance, figure 9 of ref. [21]), which is of the order of minutes in the atmosphere. Conversely, the correlation timescale of accelerations is controlled by the smallest timescale[21–23] (the dissipation timescale in figure 7 of ref. [21]). This is estimated to be only a fraction of a second, which is much smaller than the time interval between successive actions. Note that the previous experimental observations can be rationalized by the combination of the power-law spectrum of turbulent velocity fluctuations in the atmosphere and the extra factor of frequency squared in the spectrum of acceleration versus velocity fluctuations[23]. Finally, the glider's experiences, represented as state–action–state–reward quadruplets, (s_t, a_t, s_{t+1}, r_t), were cumulatively collected (over 15 days) into a set E using explorative behavioural policies. Learning is monitored by bootstrapping the standard deviation of the Q values from E (Fig. 2a), calculated through value iteration methods (see Methods).

The navigational strategy derived at the end of the training period is presented in Fig. 2b, which shows the actions deemed optimal for the 45 possible states. The rows corresponding to $\omega = 0$ resemble the Reichmann rules[24]—a set of simple heuristics for soaring, which suggest a decrease (increase) in bank angle when the climb rate increases (decreases). Our strategy also gives a prescription for bank: for instance, when a_z and ω are both positive (top row in Fig. 2b)—that is, in a situation when better lift is available diagonal to the glider's heading—it is advantageous not to bank to the extreme but rather to maintain an intermediate value between −30° and −15°. Importantly, the learned leftward (rightward) bias in bank angle on encountering a positive (negative) torque validates our estimation procedure for ω.

In Fig. 3a, we show a sample trajectory of a glider that used the navigational strategy in the field to remain aloft for about 12 min while spiralling to the height of low-lying clouds (see also Extended Data Fig. 1). On a day with strong atmospheric convection, the time spent aloft is limited only by visibility and the receiver's range as the glider soars higher or is constantly pushed away by the wind. A significant improvement in median climb rate of 0.35 m s^{-1} was measured in the field by performing repeated 3-min trials over 5 days (Fig. 3b, Mann–Whitney $U = 429$, $n_{\text{control}} = 37$, $n_{\text{strategy}} = 49$, $P < 10^{-4}$ two-sided). Notably, this value reflects a general improvement in performance averaged across widely variable conditions without controlling for the availability of nearby thermals.

To examine possible advantages of larger gliders due to improved estimation of torque, we further analysed soaring performance for different wingspans (l). Although the naive expectation is that the signal-to-noise ratio in the estimate of ω scales linearly with l, we show that the effects of atmospheric turbulence lead to a much weaker $l^{1/6}$ scaling (see Methods). Because testing our prediction would require a series of gliders with different wingspans, we turned to numerical simulations of the convective boundary layer, adapted to reflect our experimental set-up (Methods). Results shown in Fig. 3c, d are consistent with the predicted scaling. Intuitively, the weak 1/6 exponent arises because the improvement in estimation of the gradient is offset by the larger turbulent eddies, which only have a sweeping effect for smaller wingspans (that is, they do not rotate the glider but translate it, which does not affect the estimate of vertical velocity differences across its wings), and contribute to velocity differences across the wings as l increases. Our calculation yields an estimate of the signal-to-noise ratio of about 4 for typical experimental values; similar arguments for a_z yield a signal-to-noise ratio of about 7. Experimental results, together with

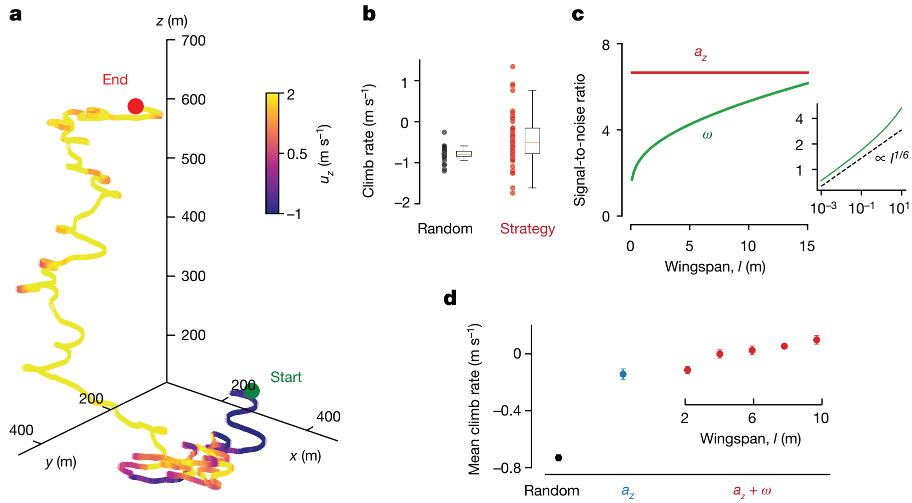


Fig. 3 | Performance of the learned strategy and its dependence on the wingspan. a, A 12-min-long trajectory of the glider executing the learned strategy for navigating thermals in the field, coloured according to the vertical ground velocity at each instant. **b**, Experimentally measured climb rate for a control random policy (black dots) is compared against the learned strategy (red dots) over repeated 3-min trials in the field. Each dot represents the average climb rate in a single trial. To restrict the range of the axis, a few outliers are not shown. The limits on the y axis are from -2 m s^{-1} to 1.5 m s^{-1}. The orange line in the box plot shows the median, the extent of the boxes marks the interquartile range and the whiskers demarcate the outliers (1.5 times the interquartile range above (below) the upper (lower) quartile). **c**, Signal-to-noise ratio for estimating ω (green) and a_z (red) as a function of wingspan. The signal-to-noise ratio for ω estimation is plotted in logarithmic scale (inset) to highlight the weak $l^{1/6}$ scaling. **d**, The mean climb rate for the learned strategy is compared for different wingspans (red filled circles) in simulations of a glider soaring in the convective boundary layer. For comparison, we show the mean climb rates for a random policy and for a strategy that uses a_z only (see Methods). Error bars represent s.e.m.

simulations and signal-to-noise ratio estimates, establish a_z and ω as robust navigational cues for thermal soaring.

The real-world intricacies of soaring impose severe constraints on the complexity of the underlying models, reflecting a fundamental trade-off between learning speed and performance. Notably, the choice of a proper reward signal was crucial to make learning feasible with the limited samples available. Although reward shaping has received some attention in the machine learning community[25], its relevance to animal behaviour remains poorly understood. We remark that our navigational strategy constitutes a set of general reactive rules, with no learning occurring during a particular thermal encounter. A soaring bird may use a model-based approach of constantly updating its estimate of the location of nearby thermals based on recent experience and visual cues. Still, the importance of vertical wind accelerations and torques for our policy suggests that they are likely to be useful for any other strategy; our methods of estimating them in a glider suggest that they should be accessible to birds as well. The hypothesis that birds use those mechanical cues while soaring can be tested in experiments.

Finally, we note that single-thermal soaring is just one face of a multifaceted question: how should a migrating bird or a cross-country glider fly among thermals over hundreds of kilometres for a quick, yet risk-averse, journey[26–28]? This calls for the development of effective methods for identifying areas of strong updraft based on mechanical and visual cues. Such methods, coupled with our current work, would pave the way to a better understanding of how birds migrate and the development of autonomous vehicles that can fly for long distances and long periods with minimal energy cost.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0533-0.

Received: 20 February 2018; Accepted: 20 July 2018;
Published online 19 September 2018.

1. Newton, I. *Migration Ecology of Soaring Birds* 1st edn (Elsevier, Amsterdam, 2008).
2. Shamoun-Baranes, J., Leshem, Y., Yom-tov, Y. & Liechti, O. Differential use of thermal convection by soaring birds over central Israel. *Condor* **105**, 208–218 (2003).
3. Weimerskirch, H., Bishop, C., Jeanniard-du-Dot, T., Prudor, A. & Sachs, G. Frigate birds track atmospheric conditions over months-long transoceanic flights. *Science* **353**, 74–78 (2016).
4. Pennycuick, C. J. Thermal soaring compared in three dissimilar tropical bird species, *Fregata magnificens*, *Pelecanus occidentals* and *Coragyps atratus*. *J. Exp. Biol.* **102**, 307–325 (1983).
5. Garrat, J. R. *The Atmospheric Boundary Layer* (Cambridge Univ. Press, Cambridge, 1994).
6. Lenschow, D. H. & Stephens, P. L. The role of thermals in the atmospheric boundary layer. *Boundary-Layer Meteorol.* **19**, 509–532 (1980).
7. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* 1st edn (MIT Press, Cambridge, 1998).
8. Tesauro, G. Temporal difference learning and TD-Gammon. *Commun. ACM* **38**, 58–68 (1995).
9. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
10. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
11. Kim, H. J., Jordan, M. I., Sastry, S. & Ng, A. in *Advances in Neural Information Processing Systems* Vol. 16 (eds Thrun, S. et al.) 799–806 (MIT Press, Cambridge, 2004).
12. Levine, S., Finn, C., Darrell, T. & Abbeel, P. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.* **17**, 1–40 (2016).
13. Allen, M. J. & Lin, V. Guidance and control of an autonomous soaring vehicle with flight test results. In *45th AIAA Aerospace Sciences Meeting and Exhibit* 2007-867 (AIAA, 2007).
14. Edwards, D. J. Implementation details and flight test results of an autonomous soaring controller. In *AIAA Guidance, Navigation and Control Conference and Exhibit* 2008-7244 (AIAA, 2008).
15. Edwards, D. J. *Autonomous Soaring: The Montague Cross Country Challenge*. PhD thesis, North Carolina State Univ. (2010).
16. Ákos, Z., Nagy, M., Leven, S. & Vicsek, T. Thermal soaring flight of birds and unmanned aerial vehicles. *Bioinspir. Biomim.* **5**, 045003 (2010).
17. Doncieux, S., Mouret, J. B. & Meyer, J.-A. Soaring behaviors in UAVs: 'animat' design methodology and current results. In *3rd US–European Competition and Workshop on Micro Air Vehicle Systems (MAV07) and European Micro Air Vehicle Conference and Flight Competition (EMAV2007)* (2007); http://www.isir.upmc.fr/files/2007ACTI734.pdf.
18. Wharington, J. & Herszberg, I. Control of a high endurance unmanned aerial vehicle. In *21st Congress of International Council of the Aeronautical Sciences* 98-3.7.1 (ICAS, 1998).
19. Chung, J. J., Lawrance, N. R. J. & Sukkarieh, S. Learning to soar: resource-constrained exploration in reinforcement learning. *Int. J. Robot. Res.* **34**, 158–172 (2015).
20. Reddy, G., Celani, A., Sejnowski, T. & Vergassola, M. Learning to soar in turbulent environments. *Proc. Natl Acad. Sci. USA* **113**, E4877–E4884 (2016).
21. Yeung, P. K. & Pope, S. B. Lagrangian statistics from direct numerical simulations of isotropic turbulence. *J. Fluid Mech.* **207**, 531–586 (1989).

22. Voth, G. A., La Porta, A., Crawford, A. M., Alexander, J. & Bodenschatz, E. Measurement of particle accelerations in fully developed turbulence. *J. Fluid Mech*. **469**, 121–160 (2002).
23. Tennekes, H. & Lumley, J. L. *A First Course in Turbulence* (MIT Press, Cambridge, 1972).
24. Reichmann, H. *Cross-Country Soaring* (Thomson Publications, Santa Monica, 1988).
25. Ng, A. Y., Harada, D. & Russell, S. J. Policy invariance under reward transformations: theory and application to reward shaping. In *Proc. 16th International Conference on Machine Learning* (eds Bratko, I. & Dzeroski, S.) 278–287 (Morgan Kaufmann, San Francisco, 1999).
26. MacCready, P. B. J. Optimum airspeed selector. *Soaring* **1958**, 10–11 (1958).
27. Horvitz, N. et al. The gliding speed of migrating birds: slow and safe or fast and risky? *Ecol. Lett.* **17**, 670–679 (2014).
28. Cochrane, J. H. MacCready theory with uncertain lift and limited altitude. *Tech. Soaring* **23**, 88–96 (1999).

Acknowledgements This work was supported by Simons Foundation grant 340106 (to M.V.) and NSF grant NCS-FO-1735004 (to T.J.S.).

Reviewer information *Nature* thanks M. Chertkov and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors were involved in designing the study and drafting the final manuscript. G.R. and J.W.N. performed the experiments and analysed the data. G.R., A.C. and M.V. contributed to the theoretical results.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0533-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0533-0.

Correspondence and requests for materials should be addressed to M.V.

METHODS

Experimental set-up. A Parkzone Radian Pro fixed-wing plane of 2-m wingspan was equipped with an on-board Pixfalcon autonomous flight controller operating on custom-modified Arduplane firmware (http://www.ardupilot.org). The instrumentation available to the flight controller includes a GPS, compass, barometer, airspeed sensor and an inertial measurement unit. Measurements from multiple instruments are combined by an extended Kalman filter (EKF) to give an estimate of relevant quantities such as the altitude z, the sink rate with respect to the ground $-u_z$, pitch ϕ, bank angle μ and the airspeed V, at a rate of 10 Hz (see Extended Data Fig. 2 for the definitions of the angles). Throughout the paper, we use $\mu > 0$ when the glider is banked to the right and $\phi > 0$ for the glider pitched with its nose above the horizontal plane. For a given desired pitch ϕ_d and desired bank angle μ_d, the controller modulates the aileron and elevator control surfaces at 400 Hz by using a proportional–integral–derivative feedback control mechanism at a user-set timescale τ (see Extended Data Table 1 for parameter values) such that:

$$\tau \frac{d\phi}{dt} = \phi_d - \phi \tag{1}$$

$$\tau \frac{d\mu}{dt} = \mu_d - \mu \tag{2}$$

The desired pitch is fixed during flight and can be used to indirectly modulate the angle of attack, α, which determines the airspeed and sink rate with respect to air of the glider ($-v_z$). Actions of increasing, decreasing or keeping the same bank angle are taken in time steps of t_a by changing μ_d such that μ increases linearly from μ_i to μ_f in time interval t_a:

$$\mu_d(t) = \mu_i + (\mu_f - \mu_i)\frac{t+\tau}{t_a} \tag{3}$$

Estimation of the vertical wind acceleration. The vertical wind acceleration is defined as:

$$a_z \equiv \frac{dw_z}{dt} = \frac{d}{dt}(u_z - v_z) \tag{4}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the velocities of the glider with respect to the ground and air respectively, and $\boldsymbol{w}$ is the wind velocity. Here, we have used the relation $\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{v}$. An estimate of $\boldsymbol{u}$ is obtained in a straightforward manner from the EKF, which combines the GPS and barometer readings to form the estimate. However, v_z is confounded by various aerodynamic effects that affect it on timescales of a few seconds (Extended Data Fig. 3). Artificial accelerations introduced by these effects impair accurate estimation of the wind acceleration and thus alter the perceived state during decision-making and learning. Two effects strongly affect variations in v_z: (1) sustained pitch oscillations with a period of a few seconds and varying amplitude, and (2) variations in angle of attack, which occur to compensate for the imbalance of lift and weight while rolling. In Supplementary Information, we present a detailed analysis of the longitudinal motions that affect the glider, summarized here for conciseness. Changes in v_z can be approximated as:

$$\Delta v_z = -V(\Delta\alpha - \Delta\phi) \tag{5}$$

where Δ denotes the deviation from their value during steady, level flight. We obtain $\Delta\phi$ directly from on-board measurements, whereas $\Delta\alpha$ can be approximated for bank angle μ as:

$$\Delta\alpha \approx (\alpha_0 - \alpha_i)\left(\frac{1}{\cos\,\mu} - 1\right) \tag{6}$$

where α_0 is the angle of attack at steady, level flight and α_i is a parameter that depends on the geometry and the angle of incidence of the wing. The constant pre-factor $(\alpha_0 - \alpha_i)$ is inferred from experiments. Measurements of u_z together with the estimate of Δv_z are now used to estimate the vertical wind velocity w_z up to a constant term, which can be ignored as it does not affect a_z. The vertical wind acceleration a_z is then obtained by taking the derivative of w_z and is further smoothed using an exponential smoothing kernel of timescale σ_a (Extended Data Fig. 4).

Estimation of vertical wind velocity gradients across the wings. Spatial gradients in the vertical wind velocity induce a roll-wise torque on the plane, which we estimate using the deviation of the measured bank angle from the expected bank angle. The total roll-wise torque on the plane has contributions from three sources: (1) the feedback control of the plane; (2) spatial gradients in the wind including turbulent fluctuations; and (3) roll-wise moments due to various aerodynamic effects. Here, we follow an empirical approach: we note that the latter two contributions perturb the evolution of the bank angle from equation (2). We can then write an effective equation

$$\frac{d\mu}{dt} = \frac{\mu_d - \mu}{\tau} + \omega(t) + \omega_{aero}(t) \tag{7}$$

where $\omega(t)$ and $\omega_{aero}(t)$ are contributions to the roll-wise angular velocity due to the wind and aerodynamic effects, respectively. We empirically find four major contributions to ω_{aero}: (1) the dihedral effect, which is a stabilizing moment due to the effects of sideslip on a dihedral wing geometry; (2) the overbanking effect, which is a destabilizing moment that occurs during turns with small radii; (3) trim effects, which create a constant moment due to asymmetric lift on the two wings; and (4) a loss of rolling moment generated by the ailerons when rolling at low airspeeds. We quantify the contributions from the four effects and model their dependence on the bank angle (see Supplementary Information for more details on modelling and calibration). An estimate of ω is then obtained as:

$$\omega = \frac{d\mu}{dt} - \frac{\mu_d - \mu}{\tau} - \omega_{aero} \tag{8}$$

Finally, an exponential smoothing kernel is applied to obtain a smoothed ω (Extended Data Fig. 5).

Design of the learning module. The navigational component of the glider is modelled as a Markov decision process, closely following the implementation used in ref. [20]. The Markovian transitions are discretized in time into intervals of size t_a. The state space consists of the possible values taken by a_z, ω and μ. To make the learning feasible within experimental constraints and to maintain interpretability, we use a simple tile coding scheme to discretize our state space: continuous values of a_z and ω are each discretized into three states (+, 0, −), partitioned by thresholds $\pm K_a$ and $\pm K_\omega$ respectively. The thresholds are set at ±0.8 times the standard deviation of a_z and ω. Because the width of the distributions of a_z and ω can vary across days, the data obtained on a particular day are normalized by the standard deviation calculated for that day. In effect, the filtration threshold to detect a signal against turbulent 'noise' is higher on days with more turbulence. The consequence is that the behaviour of the learned strategy could change across days, adapting to the recent statistics of the environment. The bank angle takes five possible values (0°, ±15°, ±30°), while the three possible actions allow for increasing, decreasing by 15° or keeping the same bank angle. In summary, we have a total of $3 \times 3 \times 5 = 45$ states in the state space and three actions in the action space.

We choose the local vertical wind acceleration a_z obtained in the next time step as the reward function. The choice of a_z as an appropriate reward signal is motivated by observations made in simulations from ref. [20]. In Supplementary Information, we show that the obtained policy using a_z as the reward function is equivalent to a policy that also maximizes the expected gain in height.

Learning the strategy in the field. Data collected in the field are split into (s, a, s', r) quadruplets containing the current state s, the current action a, the next state s' and the obtained reward r, which are pooled to obtain the transition matrix $T(s' \mid s, a)$ and reward function $R(s, a)$. Value iteration methods are used to estimate the Q values from T and R. The learning process is offline and off-policy; specifically, we begin training with a 'random' policy that takes the three possible actions with equal probability irrespective of the current state. This behavioural policy was used for 12 out of the 15 days of training. For the other days, a softmax policy[7] with 'temperature' parameter set to 0.3 was used. For softmax training, the Q values were first estimated from the data obtained in the previous days and then normalized by the difference between the maximum and minimum Q values over the three possible actions at a particular state, as described in ref. [20].

Using a fixed, random policy as our behavioural policy has the disadvantage that it slows learning, as state–action pairs that rarely appear in the final policy are still sampled. On the other hand, it has the benefit that calibrating the parameters necessary for the unbiased measurement of a_z and ω (see Supplementary Information) is performed simultaneously with learning, which considerably reduces the number of days required in the field. Importantly, offline learning permits us to continuously monitor the variance of the estimated Q values by bootstrapping from the set E of accumulated (s, a, s', r) quadruplets up to a particular point. Specifically, $|E|$ samples are drawn with replacement from E, and Q values are obtained for each state–action pair by value iteration. The steps are repeated and the average of the bootstrapped standard deviations in Q over all the state–action pairs is used as a measure of learning progress, as shown in Fig. 2a.

We expect certain symmetries in the transition matrix and the reward function, which we exploit to expedite our learning process. Particularly, we note that the Markov decision process is invariant to an inversion of sign in the bank angle $\mu \to -\mu$. This transforms a state as $(a_z, \omega, \mu) \to (a_z, -\omega, -\mu)$ and inverts the action

from that of increasing the bank angle to decreasing the bank angle and vice versa. We symmetrize T and R as

$$T^{\text{sym}}=\frac{T^{+}+T^{-}}{2} \tag{9}$$

$$R^{\text{sym}}=\frac{R^{+}+R^{-}}{2} \tag{10}$$

where + and − denote the obtained values and those computed by applying the inverting transformation respectively. Finally, T^{sym} and R^{sym} are used to obtain a symmetrized Q function, which results in a symmetric policy as shown in Fig. 2b. To conveniently obtain the policy that uses only a_z (Fig. 3d), the above procedure is repeated with the threshold for ω (K_ω) set to infinity.

Testing the performance of the learned policy in the field. To obtain the data shown in Fig. 3b, the glider is first sent autonomously to an arbitrary but fixed location 250 m above ground level. The learned policy for thermals is then turned on, and the mean climb rate (that is, the total height gained divided by the total time) is measured over a 3-min interval. To obtain the control data, the glider instead follows a random policy, which takes the three possible actions with equal probability. Trials in which we observe little to no atmospheric convection are filtered out by imposing a threshold on the standard deviation of the vertical wind velocity over the 3-min trial. In Extended Data Fig. 6, we show the distribution of the standard deviation in w_z collected from about 240 3-min trials over 9 days. Trials below the threshold chosen as the 25th percentile mark (red dashed line) are not used for our analysis.

Testing performance for different wingspans in simulations. Soaring performance is analysed in simulations similar to those developed in ref. [20] and adapted to reflect the constraints faced by our glider and the environments typically observed in the field.

The atmospheric model consists of two components: (1) a kinematic model of turbulence that reproduces the statistics of wind velocity fluctuations in the convective atmospheric boundary layer; and (2) the positions, sizes and strengths of updrafts and downdrafts. The temporal and spatial statistics of the generated velocity field satisfy the Kolmogorov and Richardson laws[29] and the mean velocity profile in the convective boundary layer[5], as described in the supplementary information of ref. [20]. Stationary updrafts and downdrafts of Gaussian shape are placed on a staggered lattice of spacing approximately 125 m on top of the fluctuating velocity field. Specifically, their contribution to the vertical wind velocity at position r is given by

$$w_z=\pm W\mathrm{e}^{-(r_{\perp}-r_{\perp}^{0})^2/(2R^2)} \tag{11}$$

where $r_{\perp}^{0}$ is the location of the centre of the up(down)draft in the horizontal plane, W is its strength and R is its radius. W is drawn from a half-normal distribution of scale 1.5 m s^{-1}, whereas the radius is drawn from a (positive) normal distribution of mean 40 m and deviation 10 m. Gaussian white noise of magnitude 0.2 m s^{-1} is added as additional measurement noise.

We assume that the glider is in mechanical equilibrium; the lift, drag and weight forces on the glider are balanced, except for centripetal forces while turning. The parameters corresponding to the lift and drag curves and the (fixed) angle of attack are set such that the airspeed is $V=8$ m s^{-1} and the sink rate is 0.9 m s^{-1} at zero bank angle, which match those measured for our glider in the field. Control over bank angle is similar to those imposed in the experiments: that is, the bank angle switches linearly between the angles 0°, ±15°, ±30° in a time interval t_a, corresponding to the time step between actions. The glider's trajectory and wind velocity readings are updated every 0.1 s. The vertical wind acceleration is derived assuming that the glider directly reads the local vertical wind velocity. The gradients in vertical wind velocity across the wings are estimated as the difference between the vertical wind velocities at the two ends of the wings. The readings are smoothed with exponential smoothing kernels; the smoothing parameters in experiments are chosen to coincide with those that yield the greatest height gain in simulations.

Estimation of noise in gradient sensing due to atmospheric turbulence. The cues a_z and ω measure the gradients in the vertical wind velocity along and perpendicular to the heading of the glider. Updrafts and downdrafts are relatively stable structures in a varying turbulent environment. Thermal detection through gradient sensing constitutes a discrimination problem of deciding whether a thermal is present or absent given the current a_z and ω. We estimate the magnitude of turbulent 'noise' that unavoidably accompanies gradient sensing. Intuitively, turbulent fluctuations in the atmospheric boundary layer (ABL) are made up of eddies of different length scales, with the largest being the size of the height of the ABL. Energy is transferred from larger, stronger eddies to smaller, weaker eddies, and eventually dissipates at the centimetre scale owing to viscosity in the bulk and owing to the boundary at the Earth's surface. In Supplementary Information, we present an explicit calculation of the signal-to-noise ratio for ω estimation, taking into account the effect of turbulent eddies on the statistics of noise. Below, we give simple scaling arguments and refer to Supplementary Information for further details.

A glider moving at an airspeed V and integrating over a timescale T averages a_z over a length VT. For V much larger than the velocity scale of the eddies, which is typically the case, the decorrelation of wind velocities is due to the glider's motion; the eddies themselves can be considered to be frozen in time. The magnitude of the spatial fluctuations across the eddy of this size scales according to the Richardson–Kolmogorov law[29] as $(VT)^{1/3}$. The mean gradient signal when going up the gradient scales as (VT); the resultant signal-to-noise ratio in a_z scales as $(VT)^{2/3}$.

Similar arguments are applicable for ω measurements. In this case, the signal-to-noise ratio has an additional dependence on the wingspan l. The dominant contribution to the noise comes from eddies of size l, whose strength scales as $l^{1/3}$. As the glider moves a distance VT, for $l \ll VT$, it traverses VT/l distinct eddies of size l. Consequently, the noise is averaged out by a factor $(VT/l)^{-1/2}$, corresponding to the VT/l independent measurements. Multiplying these two factors, the averaged noise is proportional to $l^{5/6}(VT)^{-1/2}$. As the mean gradient (that is, the signal) is approximately l, the signal-to-noise ratio is then proportional to $l^{1/6}(VT)^{1/2}$.

From the above arguments and dimensional considerations, we get order-of-magnitude estimates of the signal-to-noise ratio (SNR) for a_z and ω estimation:

$$\mathrm{SNR}_{a_z}\propto\frac{WV^{2/3}T^{2/3}L^{1/3}}{wR} \tag{12}$$

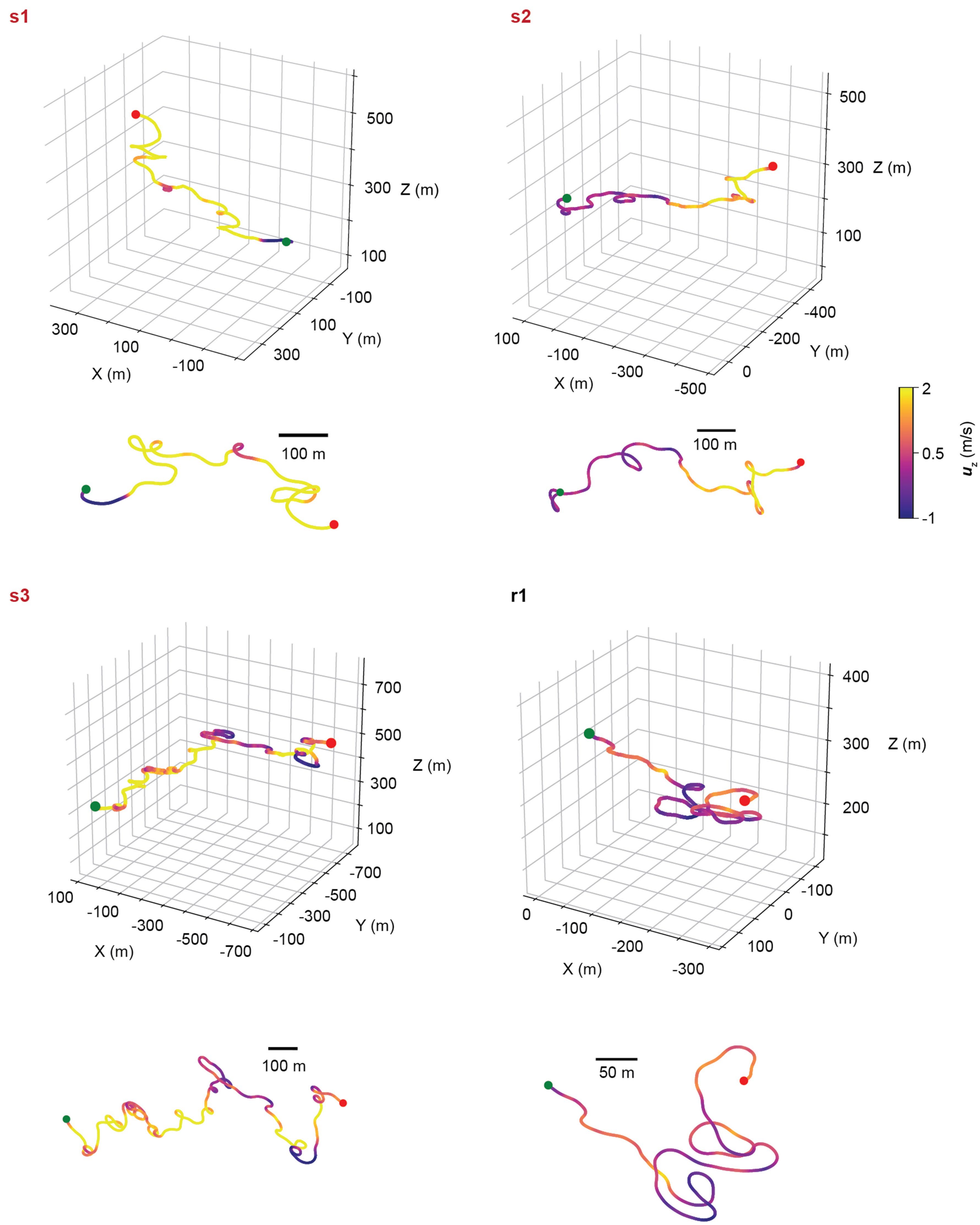
$$\mathrm{SNR}_{\omega}\propto\frac{WV^{1/2}T^{1/2}l^{1/6}L^{1/3}}{wR} \tag{13}$$

where W is the strength of the thermal, R is its radius, w is the magnitude of turbulent vertical wind velocity fluctuations and L is the length scale of the ABL. For the signal-to-noise ratio estimates presented in the text, we use $W=2$ m s^{-1}, $R=50$ m, $l=2$ m, $V=8$ m s^{-1}, $T=3$ s, $L=1$ km. The values of V and T correspond to the airspeed of the glider in experiments and the timescale between actions during learning respectively.

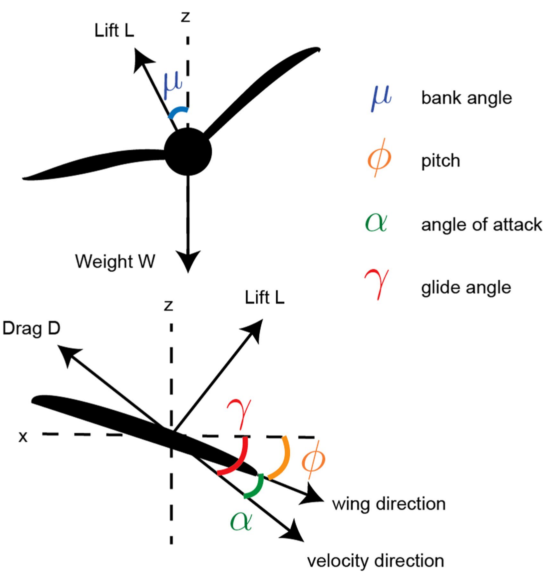
Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

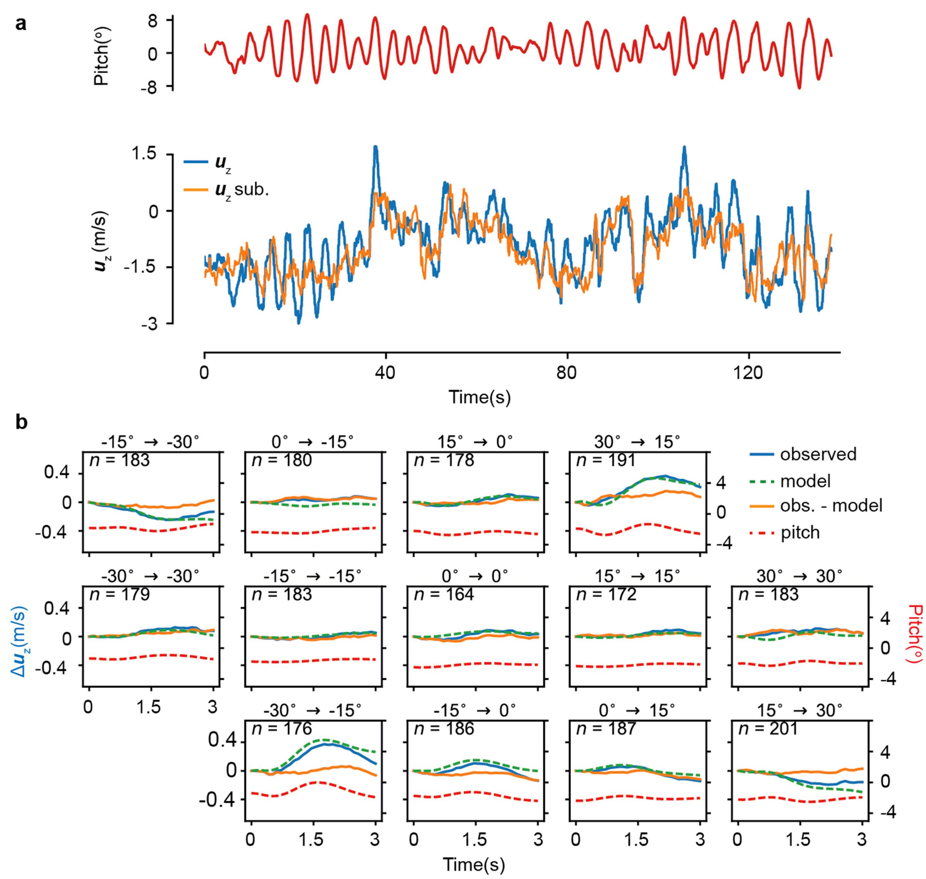
29. Frisch, U. *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge Univ. Press, Cambridge, 1995).



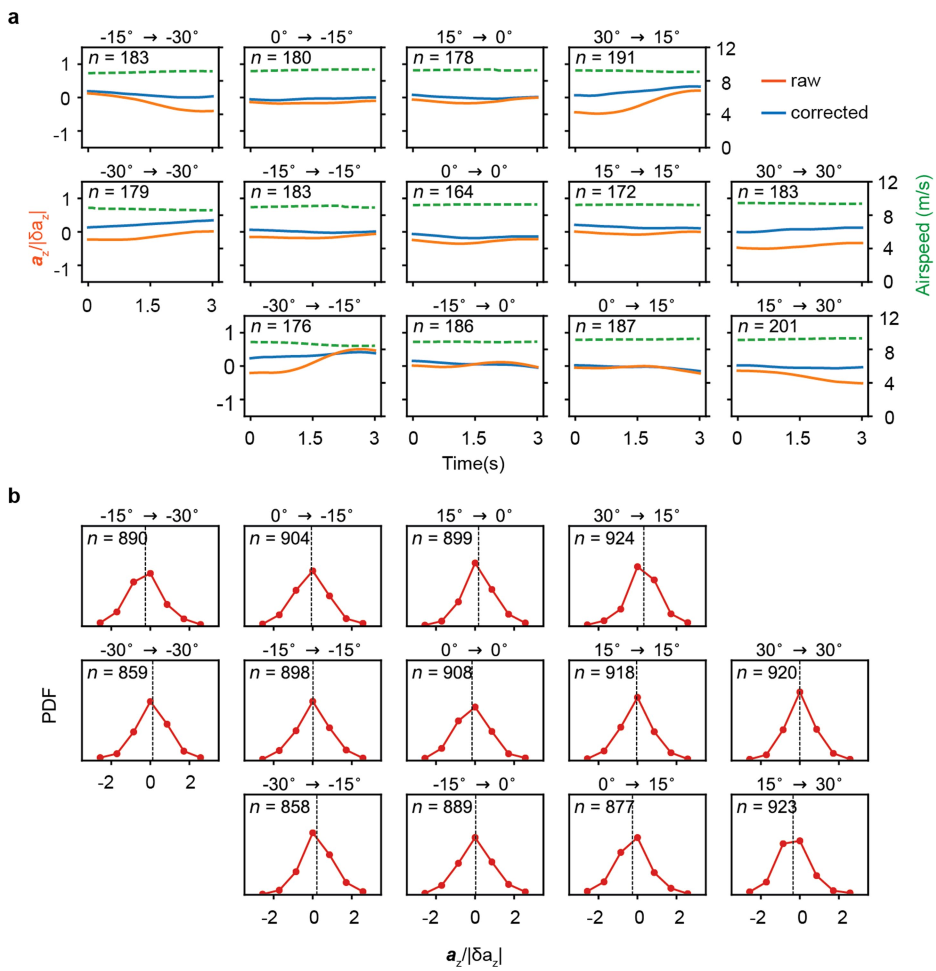
Extended Data Fig. 1 | Sample trajectories obtained in the field. The three-dimensional view and top view are shown of the glider's trajectory as it executes the learned strategy for thermals (labelled 's') or a random policy that takes actions with equal probability (labelled 'r'). The trajectories are coloured according to the instantaneous vertical ground velocity u_z. The green (red) dot shows the start (end) point of the trajectory. Trajectories s1, s2 and r1 last for 3 min each, whereas s3 lasts for about 8 min.



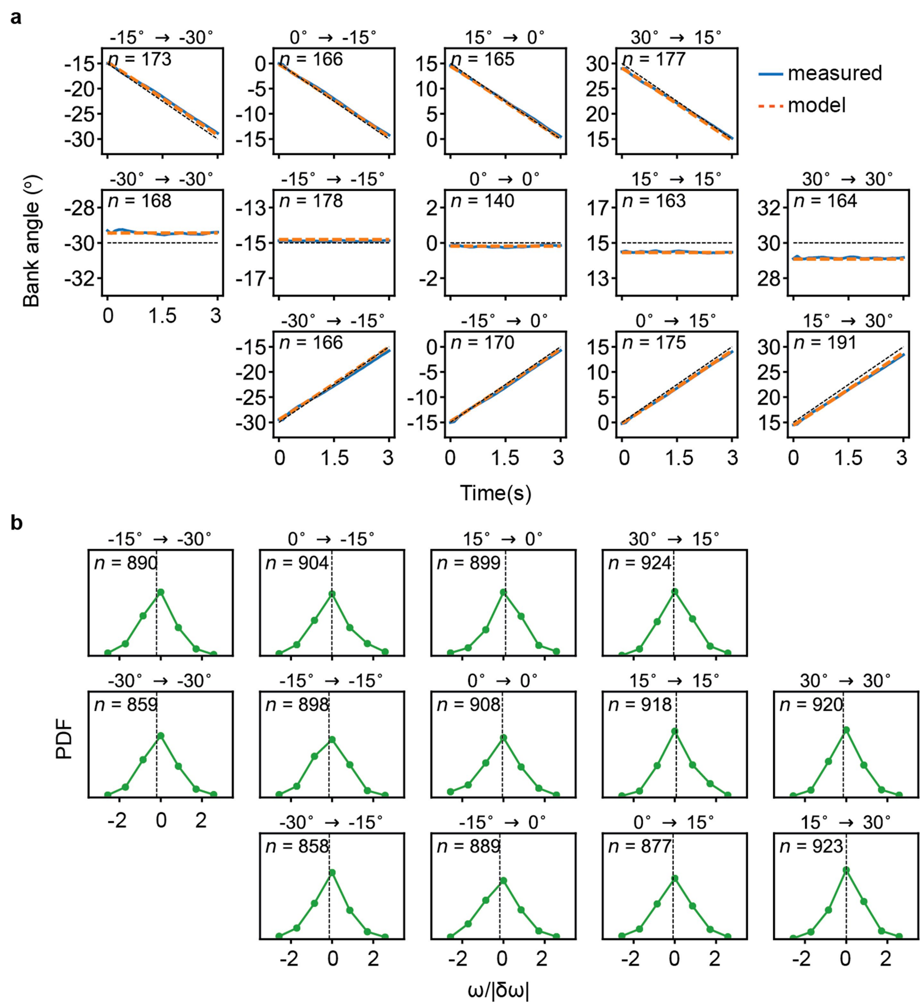
Extended Data Fig. 2 | Force–body diagram of a glider. The forces on a glider and the definitions of the various angles that determine the glider's motion.



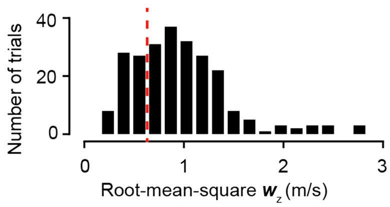
Extended Data Fig. 3 | Modelling the longitudinal motion of the glider. **a**, Sample trajectory of a glider's pitch and its vertical velocity with respect to ground (u_z) in a case in which the feedback control over the pitch is reduced in order to exaggerate the pitch oscillations. The blue line shows the measured u_z, and the orange line is u_z obtained after subtracting the contributions from longitudinal motions of the glider (see Supplementary Information). **b**, The blue line shows the average change in u_z when a particular action is taken (labelled above each panel), averaged over n 3-s intervals. The 13 panels correspond to the 13 possible bank angle changes from the angles 0°, ±15° and ±30° by increasing, decreasing the bank angle by 15° or keeping the same angle. The green dashed line shows the prediction from the model whereas the orange line is the estimated w_z. The axis on the right shows the averaged pitch (red dashed line).



Extended Data Fig. 4 | The estimated vertical wind acceleration is unbiased after accounting for the glider's longitudinal motion. **a**, The averaged vertical wind acceleration a_z in units of its standard deviation. a_z, plotted as in Extended Data Fig. 3b, is shown in orange with (blue line) and without (orange line) accounting for the glider's longitudinal motions. The axis on the right shows the airspeed (green dashed line). **b**, Probability density functions (PDFs) of a_z for the different bank angle changes. The black dashed line shows the median.



Extended Data Fig. 5 | The estimated roll-wise torque is unbiased after accounting for the effects of feedback control and glider aerodynamics. **a**, The averaged evolution of the bank angle shown as in Extended Data Fig. 3b. The blue line shows the measured bank angle and the dashed orange line shows the best-fit line obtained from simultaneously fitting the 13 blue curves to the prediction (see Supplementary Information). **b**, PDFs of the roll-wise torque ω (in units of its standard deviation) for the different bank angle changes. The black dashed line shows the median value.



Extended Data Fig. 6 | The distribution of the strength of vertical currents observed in the field. The root-mean-square vertical wind velocity measured in the field is pooled from about 240 3-min trials collected over 9 days. The dashed red line shows the threshold criterion imposed when measuring the performance of the strategy in the field (see Methods).

Extended Data Table 1 | Parameter values

Label	Description	Value
l	Wingspan of glider used in experiments	2m
φ_d	Desired pitch	-2°
τ	Feedback control time scale	0.45s
t_a	Interval between actions (learning)	3s
t_a	Interval between actions (soaring)	1.5s
$\alpha_0 - \alpha_i$	Net angle of attack (see eq. 6)	14°
V	Airspeed (typical)	6 to 8 m/s
T_{dih}	Dihedral effect timescale (typical)	14 to 30 s
T_{ob}	Overbanking effect timescale (typical)	< -20s
b	Trim bias (typical)	-2 to +2°/s
T_{roll}	Opposing roll timescale (typical)	1.5 to 3 s
$\pm K_a$, $\pm K_w$	Thresholds for $\boldsymbol{a}_z$ and ω state estimation	0.8 x std. dev
σ_a, σ_a'	Exponential smoothing timescales for $\boldsymbol{a}_z$	$8t_a/3$, $2t_a/3$
σ_w, σ_w'	Exponential smoothing timescales for ω	t_a, $t_a/4$
γ	Discount factor for RL implementation	0.8

LETTER

https://doi.org/10.1038/s41586-018-0592-2

Electronic noise due to temperature differences in atomic-scale junctions

Ofir Shein Lumbroso[1], Lena Simine[2,3], Abraham Nitzan[4,5], Dvira Segal[2] & Oren Tal[1]*

Since the discovery a century ago[1–3] of electronic thermal noise and shot noise, these forms of fundamental noise have had an enormous impact on science and technology research and applications. They can be used to probe quantum effects and thermodynamic quantities[4–11], but they are also regarded as undesirable in electronic devices because they obscure the target signal. Electronic thermal noise is generated at equilibrium at finite (non-zero) temperature, whereas electronic shot noise is a non-equilibrium current noise that is generated by partial transmission and reflection (partition) of the incoming electrons[8]. Until now, shot noise has been stimulated by a voltage, either applied directly[8] or activated by radiation[12,13]. Here we report measurements of a fundamental electronic noise that is generated by temperature differences across nanoscale conductors, which we term 'delta-T noise'. We experimentally demonstrate this noise in atomic and molecular junctions, and analyse it theoretically using the Landauer formalism[8,14]. Our findings show that delta-T noise is distinct from thermal noise and voltage-activated shot noise[8]. Like thermal noise, it has a purely thermal origin, but delta-T noise is generated only out of equilibrium. Delta-T noise and standard shot noise have the same partition origin, but are activated by different stimuli. We infer that delta-T noise in combination with thermal noise can be used to detect temperature differences across nanoscale conductors without the need to fabricate sophisticated local probes. Thus it can greatly facilitate the study of heat transport at the nanoscale. In the context of modern electronics, temperature differences are often generated unintentionally across electronic components. Taking into account the contribution of delta-T noise in these cases is likely to be essential for the design of efficient nanoscale electronics at the quantum limit.

At non-zero temperature, the thermal motion of electrons leads to temporal current fluctuations referred to as the thermal (Johnson–Nyquist) noise S_{TN}, even at zero net current, in equilibrium conditions[2,3]. This noise depends on the conductance G ($G=1/R$, where R is the resistance) and temperature T in a straightforward manner[8], $S_{\mathrm{TN}}=4k_{\mathrm{B}}TG$, where k_{B} is Boltzmann's constant. Thermal noise can be used as a primary thermometer, and it does not depend on the conductor's shape, material type or the details of the transport mechanism[4,15]. When current is generated across a conductor, electrons can either be transmitted through the conductor or backscatter, leading to non-equilibrium temporal current fluctuations called electronic shot noise. This noise has been extensively used in the study of electronic transport in quantum conductors, including the analysis of quasiparticles' charge, electronic spin transport, and interacting many-body systems[5,6,9,10]. Shot noise measurements also provide unique information about electronic transport at the miniaturization limit of electronic conductors, namely across atomic and molecular junctions[7,16–21]. These junctions are composed of individual atoms or molecules suspended between two electrodes. The conductance of such quantum coherent conductors is described by the Landauer formalism as the sum of contributions from several transmission channels[8], $G=G_0\sum_i \tau_i$. Here τ_i is the transmission probability of the ith channel, and it can take any value between zero (closed channel) to one (fully open channel). G_0 is the quantum of conductance; $G_0\approx(12.9\ \mathrm{k\Omega})^{-1}$. In the Landauer framework, the current noise in spin-degenerate quantum conductors, including both thermal and shot noises, can be described as[7,8]

$$S_I=4k_{\mathrm{B}}TG_0\sum_i \tau_i^2+2eV\coth\left(\frac{eV}{2k_{\mathrm{B}}T}\right)G_0\sum_i \tau_i(1-\tau_i) \qquad (1)$$

where e and V are the electron charge and the applied voltage across the conductor, respectively. At zero applied voltage, the contribution of shot noise is nullified (that is, for $eV\ll k_{\mathrm{B}}T$ the second term collapses to $4k_{\mathrm{B}}TG_0\sum_i \tau_i(1-\tau_i)$) and equation (1) reduces to the thermal noise.

When a temperature difference ΔT, instead of a voltage difference, is applied across the conductor, a new approximate expression for the current noise can be derived based on the Landauer formalism:

$$S_I\approx 4k_{\mathrm{B}}\bar{T}G_0\sum_i \tau_i+\left[\frac{k_{\mathrm{B}}(\Delta T)^2}{\bar{T}}\left(\frac{\pi^2}{9}-\frac{2}{3}\right)\right]G_0\sum_i \tau_i(1-\tau_i) \qquad (2)$$

Here, $\bar{T}$ is the arithmetic average of T_{h} and T_{c}, the temperatures at the hot and cold sides of the conductor (Fig. 1a), and $\Delta T=T_{\mathrm{h}}-T_{\mathrm{c}}$. An expression for the noise generated by temperature difference has been previously derived for diffusive conductors[22]. The full derivation of equation (2), including additional terms and more general expressions, appears in Supplementary Information. The first term corresponds to the thermal noise. However, when a temperature difference is applied, this term depends on the average temperature across the conductor. Remarkably, a new noise contribution (the second term), which we denote as delta-T noise, is generated as a result of the temperature difference across the conductor. In contrast to standard voltage-activated shot noise, delta-T noise has a pure thermal origin. Yet, similarly to standard shot noise, it depends on the factor $\sum_i \tau_i(1-\tau_i)$ despite the absence of a voltage gradient across the conductor. This dependence is a signature of electronic partition noise[8], namely, noise that is activated by the partial transmission and backscattering of transporting electrons. For delta-T noise, non-equilibrium conditions are introduced by a temperature difference and the partition noise is activated even in the ideal case of zero net charge current owing to opposite and equal currents above and below the chemical potential (Fig. 1b). Thus, delta-T noise can be viewed as shot noise that is generated by temperature difference.

To experimentally demonstrate the effect of temperature difference on the noise generated in a quantum conductor, we study molecular junctions based on hydrogen molecules introduced between two atomically sharp gold electrodes[23,24]. We use the break junction technique[25] (Fig. 1a and Methods) to form an ensemble of molecular junctions with different local structure and hence different conductance values (Extended Data Fig. 1). In contrast to bare gold atomic junctions, the hydrogen-based molecular junctions provide a wide span of conductance values below $1G_0$. Shot noise measurements indicate that below $1G_0$ the conductance of the formed molecular junctions is typically

[1]Department of Chemical and Biological Physics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Chemistry, University of Toronto, Toronto, Ontario, Canada. [3]Department of Chemistry, Rice University, Houston, TX, USA. [4]Department of Chemistry, University of Pennsylvania, Philadelphia, PA, USA. [5]School of Chemistry, Tel Aviv University, Tel Aviv, Israel. *e-mail: oren.tal@weizmann.ac.il

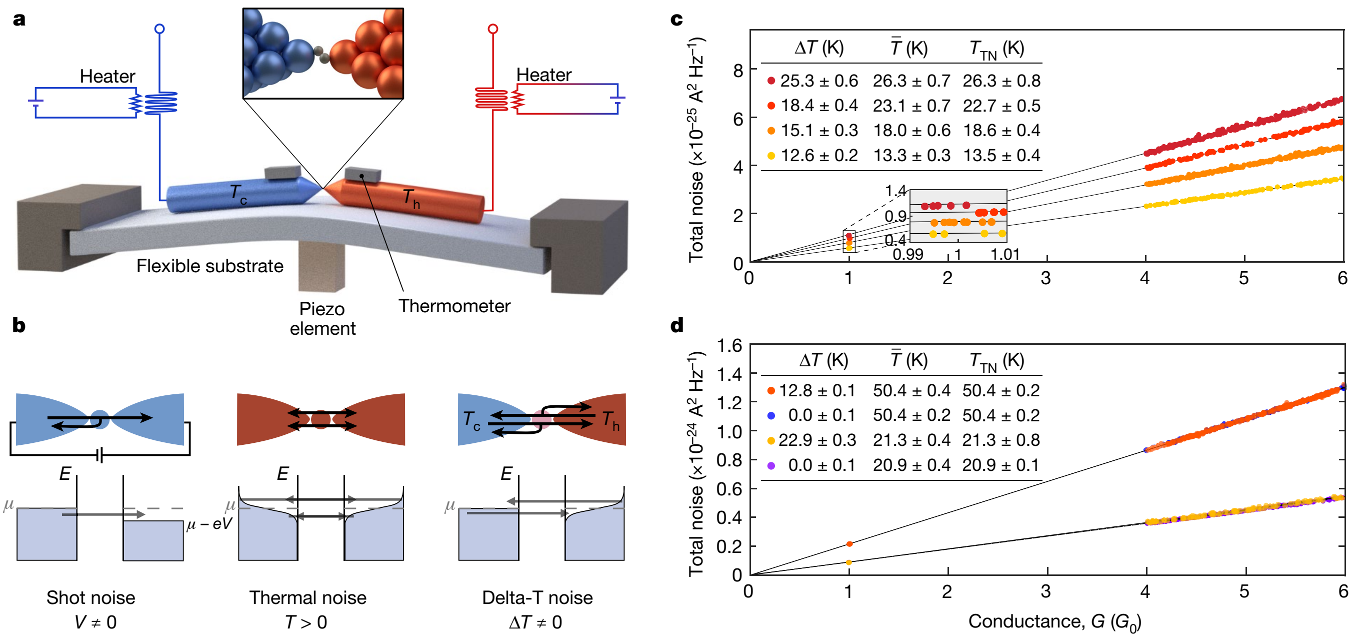


Fig. 1 | Experimental setup, noise contributions and measured total noise at a finite temperature difference. a, Schematic of the break junction setup, and the Au/H_2 junction. **b**, Illustration of the standard shot noise, thermal noise and delta-T noise generated in atomic-scale junctions. For simplicity, in the left and right schematics we assume $T = 0$ and $T_c = 0$, respectively. μ is the chemical potential; V is the applied voltage across the junction; e is the electron charge. **c**, Total noise as a function of conductance measured in Au/H_2 junctions at different temperatures T_h and T_c at the opposite sides of the junctions. The linear dependence of the noise on the conductance is expected when the total noise is dominated by the first term in equation (2) (thermal noise), while the second term (delta-T noise) is suppressed. This situation is expected at $1G_0$ and $>4G_0$. The inset tables present the temperature difference ΔT and average temperature $\bar{T}$ determined by the thermometers at the opposite sides of the junctions, and the thermal noise temperature T_{TN}, which is extracted from the slope of the total noise. For a given conductance, the thermal noise is exclusively determined by the average temperature of the hot and cold electrodes. **d**, Four sets of total noise data as in **c**. Each pair of datasets is taken at a similar $\bar{T}$, but one set is measured at $\Delta T \neq 0$ and the other at $\Delta T = 0$. The data presented illustrate that a comparable thermal noise is generated at different ΔT values, as long as $\bar{T}$ is identical. The error bars of the total noise data, corresponding to the systematic errors in our measurements, are smaller than the diameter of the symbols. Nine measurement sets at different ΔT were collected on three different samples with similar results. In each presented set, 91–301 junctions were realized and measured.

governed by a single dominant transmission channel, with minor contributions from secondary channels (see Methods and Extended Data Fig. 4). The transmission probabilities of these channels can be varied, for example, by adjusting the separation between the electrodes in sub-ångström resolution[25]. A temperature gradient across the junction was applied by an asymmetric heating of the junction's electrodes above a base 4.2 K. The temperature difference across the junction was monitored by two thermometers located at opposite sides of the junction (Fig. 1a). To determine the temperature at the nanoscale vicinity of the junction, the thermometers were calibrated using the thermal noise generated in the junction, when no temperature difference was applied (see Methods).

Equation (2) represents current noise due to temperature difference across a quantum coherent conductor as an additive combination of a standard thermal noise, yet proportional to the average temperature, and a new contribution associated with thermal difference. To test the validity of the first term in equation (2), we consider cases where a temperature gradient is applied across the examined junctions, while the second term (delta-T noise) is suppressed. Practically, this situation can be met in two ways. When the conductance of the studied junction is dominated by a single channel with transmission probability close to one ($\tau \approx 1$), the second term is expected to be very small. This condition is indeed achieved in some junction realizations, as indicated by shot noise measurements (Extended Data Fig. 4). Furthermore, the relative contribution of the second term with respect to the first one depends on $(\Delta T/\bar{T})^2$ and the Fano factor $F = \sum_i \tau_i(1-\tau_i)/\sum_i \tau_i$. The Fano factor can be determined by shot noise measurements on similar junctions. We found that if the junction is squeezed to form a multi-atomic gold contact[7] with a conductance above $4G_0$, the maximal contribution of the second term in equation (2) is less than 5% of the magnitude of the first term in the examined conditions, and typically around 3% (see Methods and Extended Data Fig. 4).

In Fig. 1c we show the measured total noise as a function of conductance (see Methods) for the studied junctions at different average junction temperature and temperature difference, determined by the calibrated thermometers at the hot and cold sides of the junction. We attribute the linear dependence of the noise on the conductance to efficient suppression of the second term in equation (2) at $1G_0$ and above $4G_0$. In these conditions, the total noise is practically reduced to the thermal noise, and the temperature associated with the thermal noise T_{TN} can be extracted from the slope of each curve. The inset table in Fig. 1c shows that $T_{TN} = \bar{T}$ within the error range, indicating that the thermal noise generated in the junction depends on the average temperature of the junction. Figure 1d presents two examples for total noise versus conductance measured at comparable average temperature of about $\bar{T} = 21$ K, as well as $\bar{T} = 50$ K. In each example, the temperature difference across the examined junction is set to be either zero ($\Delta T = 0$) or finite ($\Delta T \neq 0$), as seen in the inset table. The data points clearly fall on top of each other, illustrating that the thermal noise is exclusively determined by the average temperature and that it is independent on the temperature difference.

We now focus on the identification of the delta-T noise, and its properties. Figure 2 presents measurements of excess noise as a function of conductance for different temperature differences and average temperatures. We examined the conductance range $0.1 < G < 1G_0$ to look for a possible $\tau(1-\tau)$ dependence of the delta-T noise. The excess noise is defined as the total noise minus the average thermal noise. The latter is obtained as presented in Fig. 1c and explained above. The sets of measurements at $\Delta T = 0$ K (Fig. 2a–c) show data around zero excess noise. In the absence of temperature difference, the total noise is governed by the thermal noise. Therefore, subtracting the average thermal

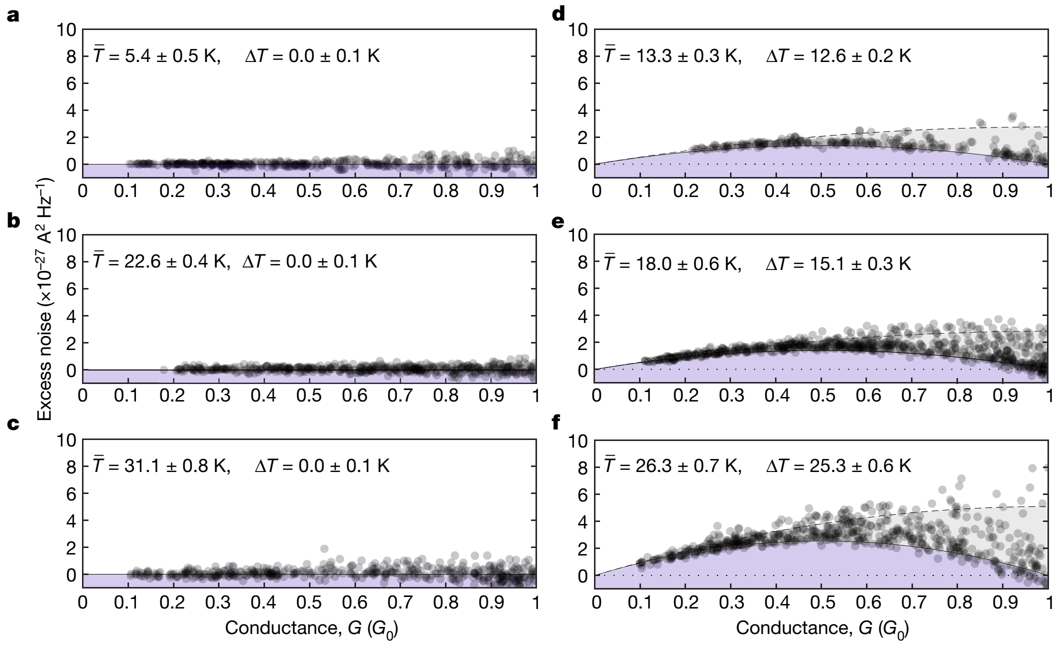


Fig. 2 | Excess noise measured at zero and finite temperature difference. **a**–**c**, Excess noise (obtained by subtracting the average thermal noise from the total measured noise) versus conductance measured in the molecular junctions examined at different temperatures at thermal equilibrium ($\Delta T = 0$). **d**–**f**, Excess noise versus conductance measured at different average temperatures and finite temperature differences across the junctions ($\Delta T \neq 0$). Calculated delta-T noise is given by the black curve for a single transmission channel, and by the dashed curve for two channels with equal transmission probabilities (non-approximated numerical calculations based on equation (S2) in Supplementary Information). The error bars, corresponding to the systematic errors in our measurements, are comparable or slightly larger than the diameter of the dark semitransparent symbols, as shown in Extended Data Fig. 7. When a temperature difference is applied across the junctions, a clear enhancement of the excess noise is observed. The measured excess noise can be described by the theoretical expression for the delta-T noise. The spread in the results is a natural outcome of additional transmission channels that open up as the conductance increases (see Extended Data Fig. 4). Eight measurement sets at different ΔT were collected on three different samples with similar results. In each presented set, 248–716 junctions were realized and measured.

noise from the total noise gives values that are scattered around zero. Remarkably, when a temperature difference is applied across the junction, an excess noise is activated (Fig. 2d–f), indicating that the origin of the measured noise is temperature difference. We note that even in the absence of applied voltage, the thermoelectric effect can generate voltage in the presence of a temperature difference[26]. This thermoelectric voltage produces shot noise that can be detected as an excess noise at a finite temperature difference. However, in the examined junctions, when no external voltage is applied, the expected shot noise due to the thermoelectric voltage is about three orders of magnitude lower than the measured excess noise in Fig. 2d–f. This is illustrated in Extended Data Fig. 6, by measuring the total thermoelectric voltage produced in our experiments and calculating the shot noise that can be generated by the highest thermoelectric voltage that was found (see Methods). We therefore conclude that the contribution of shot noise due to the generated thermoelectric voltage (in the absence of applied voltage) is negligible with respect to the measured excess noise at a finite temperature difference, and cannot explain its origin.

In Fig. 2d–f, the dependence of the low-lying noise data on the conductance is well described by the solid curve, which provides the calculated delta-T noise, assuming a single transmission channel. In fact, the majority of the data points accumulate in the vicinity of this curve, indicating the activation of delta-T noise in junctions with a dominant conductance contribution from a single transmission channel. As the conductance increases, the spread of the measured noise towards higher values increases as well. This characteristic trend is captured by the dashed line that gives the calculated delta-T noise for junctions with two channels of equal transmission probabilities ($\tau_1 = \tau_2$ and $\tau_1 + \tau_2 = G/G_0$). Shot noise measurements and numerical channel analysis (see Methods and Extended Data Fig. 4) indicate that most of the examined molecular junctions are characterized by transport via a dominant channel, yet some junctions can have a large and even comparable conductance contribution from a second channel. Additional channels, beyond the first two, usually have either a minor contribution or no contribution. From this channel analysis, delta-T noise is expected to yield excess noise data that are mainly located within the grey region, as we indeed observe. Thus, the characteristics of the measured excess noise fit the expected behaviour of delta-T noise in the examined junctions. Similar measurements were performed on bare gold atomic junctions, yet with a very limited span of conductance below $1G_0$ (Extended Data Fig. 8). Using equation (2), we can extract the Fano factor from the excess noise that is generated by temperature difference. Extended Data Fig. 9 shows that the Fano factor distribution acquired in this way and plotted versus conductance is similar to the one obtained by voltage-activated shot noise measurements (Extended Data Fig. 4). This comparison further demonstrates that the excess noise at finite temperature difference is the delta-T noise described by the second term of equation (2).

The quadratic dependence of the delta-T noise on temperature difference is a distinctive fingerprint of this noise. To check whether the detected excess noise shows the expected dependence on temperature difference, we normalize the measured excess noise, based on equation (2), and plot it with respect to ΔT in Fig. 3. The normalization is given in the caption of Fig. 3, assuming a single channel. The data spread for each ΔT is asymmetric and can be described by a generalized extreme value distribution (inset to Fig. 3). The red rectangles in Fig. 3 give the most probable values of the normalized excess noise, which are determined by the peak of the fit to the data distribution (see the inset of Fig. 3) for each temperature difference. The dashed curve in the main panel of Fig. 3 depicts the quadratic dependence of the delta-T noise on the temperature difference for a single channel scenario, and it fits very well the most probable normalized excess noise. The upward spread of the data (transparent circles) is attributed to the presence of junctions with more than one transmission channel, since the noise at

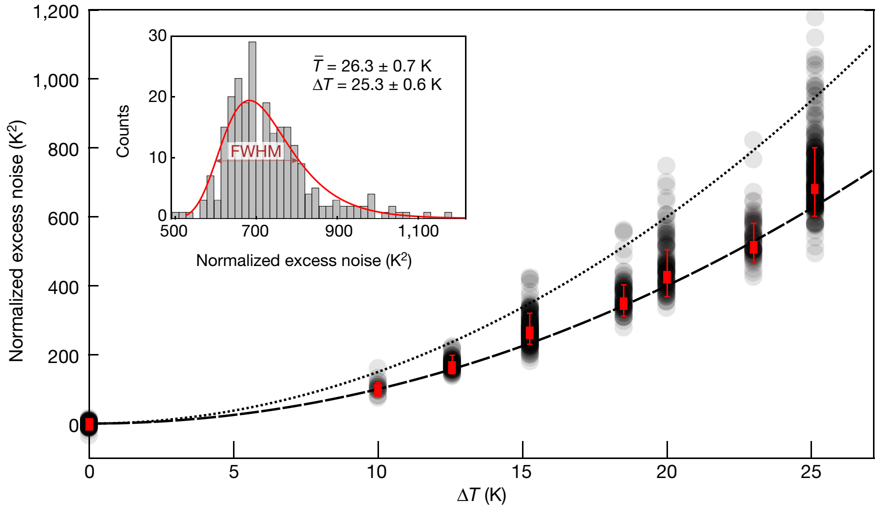


Fig. 3 | Excess noise dependence on the applied temperature difference. Normalized excess noise as a function of ΔT (black semitransparent symbols). To test the possible $(\Delta T)^2$ dependence expected for the delta-T noise, the measured excess noise between $0.1G_0$ and $0.5G_0$ was normalized (by dividing by $k_B G_0 \tau(1-\tau)(\pi^2/9-2/3)/T_{TN}$ on the basis of the second term in equation (2)), assuming $\bar{T}=T_{TN}$. The dashed line shows the calculated normalized delta-T noise for the case of a single transmission channel, and the dotted line illustrates the calculated normalized delta-T noise for the case of two channels with equal transmission probabilities. The most probable normalized excess noise (red rectangles) shows a clear $(\Delta T)^2$ dependence, as expected for delta-T noise that is generated by a single channel. The inset shows the distribution of the normalized excess noise for $\Delta T = 25.3 \pm 0.6$ K. The most probable normalized excess noise is determined by the maximum of a fitted generalized extreme value distribution that captures the asymmetric distribution of the data. Error bars are determined by the full-width at half-maximum (FWHM) as illustrated in the inset. The measured excess noise is normalized assuming a single channel. As a result, the spread of the data (black transparent circles) is artificially increased towards higher values.

a given conductance is larger for a higher number of partially open channels, as a result of its dependence on $\sum_i \tau_i(1-\tau_i)$. Nevertheless, most of the data falls below the dotted curve, which shows the delta-T noise dependence on $(\Delta T)^2$ for the case of two channels with similar transmission probabilities. Observing this dependence provides a complementary indication that the measured noise behaves as expected for the delta-T noise.

To conclude, our experimental findings, supported by a theoretical derivation, demonstrate that electronic noise emerges in the presence of a temperature difference across quantum conductors. We term this noise contribution as the delta-T noise and show that it possesses a peculiar combination of characteristics that makes it distinct from the standard thermal noise and voltage-activated shot noise. Beyond the fundamental interest in the observation and characterization of a temperature-difference-based form of partition noise, the delta-T noise can be used (in combination with thermal noise) as a probe for temperature differences. This ability is particularly interesting for nanoscale systems since fabricating physical probes that measure local temperature at this scale is extremely challenging. In contrast to physical sensors, the delta-T noise is a versatile probe, which is not limited to a specific temperature range and can be applied to conductors of different sizes, down to the atomic scale. Delta-T noise measurements can be performed without particular design limitations and can be implemented in a variety of setups, including scanning probe microscopes, nanoscale devices and even in embedded systems, which are less accessible to temperature sensing. This flexibility makes the delta-T noise an attractive tool for the study of heat management, including thermoelectricity, heat pumping and heat dissipation, which are important processes in the context of energy saving and sustainable energy production. Finally, temperature gradients are often unintentionally produced in electronic circuits. Thus, in the process of electronics miniaturization towards the quantum limit, the delta-T noise could become a performance-limiting factor that should be suppressed by minimizing temperature gradients.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0592-2.

Received: 7 May 2018; Accepted: 20 August 2018;
Published online 10 October 2018.

1. Schottky, W. Über spontane Stromschwankungen in verschiedenen Elektrizitätsleitern. *Ann. Phys.* **362**, 541–567 (1918).
2. Johnson, J. B. Thermal agitation of electricity in conductors. *Nature* **119**, 50–51 (1927).
3. Nyquist, H. Thermal agitation of electric charge in conductors. *Phys. Rev.* **32**, 110–113 (1928).
4. White, D. R. et al. The status of Johnson noise thermometry. *Metrologia* **33**, 325–335 (1996).
5. Saminadayar, L., Glattli, D. C., Jin, Y. & Etienne, B. Observation of the e/3 fractionally charged Laughlin quasiparticle. *Phys. Rev. Lett.* **79**, 2526 (1997).
6. de Picciotto, R. et al. Direct observation of a fractional charge. *Nature* **389**, 162–164 (1997).
7. van den Brom, H. E. & van Ruitenbeek, J. M. Quantum suppression of shot noise in atom-size metallic contacts. *Phys. Rev. Lett.* **82**, 1526–1529 (1999).
8. Blanter, Y. M. & Büttiker, M. Shot noise in mesoscopic conductors. *Phys. Rep.* **336**, 1–66 (2000).
9. Roche, P. et al. Fano factor reduction on the 0.7 conductance structure of a ballistic one-dimensional wire. *Phys. Rev. Lett.* **93**, 116602 (2004).
10. Delattre, T. et al. Noisy Kondo impurities. *Nat. Phys.* **5**, 208–212 (2009).
11. Jezouin, S. et al. Controlling charge quantization with quantum fluctuations. *Nature* **536**, 58–62 (2016).
12. Schoelkopf, R. J., Kozhevnikov, A. A., Prober, D. E. & Rooks, M. J. Observation of 'photon-assisted' shot noise in a phase-coherent conductor. *Phys. Rev. Lett.* **80**, 2437 (1998).
13. Reydellet, L.-H., Roche, P., Glattli, D. C., Etienne, B. & Jin, Y. Quantum partition noise of photon-created electron-hole pairs. *Phys. Rev. Lett.* **90**, 176803 (2003).
14. Büttiker, M., Imry, Y., Landauer, R. & Pinhas, S. Generalized many-channel conductance formula with application to small rings. *Phys. Rev. B* **31**, 6207 (1985).
15. Giazotto, F., Heikkilä, T. T., Luukanen, A., Savin, A. M. & Pekola, J. P. Opportunities for mesoscopics in thermometry and refrigeration: physics and applications. *Rev. Mod. Phys.* **78**, 217 (2006).
16. Djukic, D. & van Ruitenbeek, J. M. Shot noise measurements on a single molecule. *Nano Lett.* **6**, 789–793 (2006).
17. Tal, O., Krieger, M., Leerink, B. & van Ruitenbeek, J. M. Electron-vibration interaction in single-molecule junctions: from contact to tunneling regimes. *Phys. Rev. Lett.* **100**, 196804 (2008).
18. Chen, R., Matt, M., Pauly, F., Cuevas, J. C. & Natelson, D. Shot noise variation within ensembles of gold atomic break junctions at room temperature. *J. Phys. Condens. Matter* **26**, 474204 (2014).
19. Burtzlaff, A., Weismann, A., Brandbyge, M. & Berndt, R. Shot noise as a probe of spin-polarized transport through single atoms. *Phys. Rev. Lett.* **114**, 016602 (2015).
20. Vardimon, R., Klionsky, M. & Tal, O. Indication of complete spin filtering in atomic-scale nickel oxide. *Nano Lett.* **15**, 3894–3898 (2015).

21. Karimi, M. A. et al. Shot noise of 1, 4-benzenedithiol single-molecule junctions. *Nano Lett.* **16**, 1803–1807 (2016).
22. Sukhorukov, E. V. & Loss, D. Noise in multiterminal diffusive conductors: Universality, nonlocality, and exchange effects. *Phys. Rev. B* **59**, 13054 (1999).
23. Csonka, S., Halbritter, A. & Mihály, G. Pulling gold nanowires with a hydrogen clamp: strong interactions of hydrogen molecules with gold nanojunctions. *Phys. Rev. B* **73**, 075405 (2006).
24. Nakazumi, T., Kaneko, S. & Kiguchi, M. Electron transport properties of Au, Ag, and Cu atomic contacts in a hydrogen environment. *J. Phys. Chem. C* **118**, 7489–7493 (2014).
25. Muller, C. J., van Ruitenbeek, J. M. & De Jongh, L. J. Experimental observation of the transition from weak link to tunnel junction. *Physica C* **191**, 485–504 (1992).
26. Reddy, P., Jang, S. Y., Segalman, R. A. & Majumdar, A. Thermoelectricity in molecular junctions. *Science* **315**, 1568–1571 (2007).

Acknowledgements O.T. appreciates the support of the Harold Perlman family, and acknowledges funding by a research grant from Dana and Yossie Hollander, the Israel Science Foundation (grant number 1089/15), and the Minerva Foundation (grant number 120865). D.S. acknowledges support from an NSERC Discovery Grant and the Canada Research Chair programme. The research of A.N. is supported by the US National Science Foundation (grant number CHE1665291), the Israel-US Binational Science Foundation, the German Research Foundation (DFG TH 820/11-1) and the University of Pennsylvania. L.S. acknowledges the Special Opportunity Graduate Travel Fellowship from the Department of Chemistry, University of Toronto.

Reviewer information *Nature* thanks E. Scheer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The project was conceived by O.T. The experiments were designed by O.S.L. and O.T., and performed by O.S.L., under the supervision of O.T. The theoretical derivation was performed by L.S., A.N. and D.S. The manuscript was written by O.S.L., D.S. and O.T. All authors contributed to the final version of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0592-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0592-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to O.T.

METHODS

Sample fabrication and the break junction technique. Our experiments were performed using a mechanically controllable break junction[25] setup located within a cryogenic chamber. The chamber is pumped to 10^{-5} mbar and then cooled down to the liquid helium temperature (4.2 K). This setup is placed in a specially designed Faraday cage to allow efficient noise measurements. The sample consists of a notched Au wire (99.99%, 0.1 mm diameter, 25 mm length, Goodfellow), which is attached to a flexible substrate (0.76 mm thick insulating Cirlex film). A three-point bending mechanism is used to bend the substrate in order to break the wire at the notch (Fig. 1a). The wire is first broken in cryogenic vacuum, to expose two clean atomically sharp tips that serve as the junction's electrodes. The breaking process is controlled by a piezoelectric element (PI P-882 PICMA), which is driven by a 24-bit NI-PCI4461 data acquisition (DAQ) card followed by a Piezomechanik SVR 150/1 piezo driver. These components provide fast and accurate control over the distance between the two tips with sub-ångström resolution. Conductance versus inter-electrode distance (conductance traces) were measured on bare Au junctions during repeated breaking and formation of the junction. Conductance histograms (for example, Extended Data Fig. 1) that provide the most probable conductance of the examined junction were constructed based on these traces to ensure that the junction exhibits the typical conductance characteristics of bare Au atomic-scale junctions[27,28].

To form molecular junctions, pure hydrogen gas (99.999%, Gas Technologies) was introduced to the junction via a stainless steel capillary that connects an external molecular source with the cryogenic environment. The flow of hydrogen was increased by increasing the hydrogen pressure up to about 10^{-2} mbar at the capillary input. The formation of Au/H_2 junctions was monitored during the insertion process by continuously recording conductance traces and producing a typical conductance histogram for Au/H_2 junctions (Extended Data Fig. 1). Following the formation of molecular junctions, the hydrogen flow was stopped. Further details concerning the characterization of molecular junctions are given in the Methods section 'Molecular junction characterization'.

Electronic measurement setup. To measure conductance traces, direct-current (d.c.) conductance is monitored while the junction is gradually broken by increasing the voltage applied on the piezoelectric element at a constant speed of 600 nm s^{-1} and a sampling rate of 100 kHz. The junction is biased with a constant voltage of 10–200 mV provided by a NI-PCI4461 DAQ card. The resulting current is amplified by a current preamplifier (SR570) and recorded by the DAQ card. Following each trace, the exposed atomic tips are pushed back into contact until the conductance reaches a value of at least $50G_0$, in order to ensure that the data consists of a statistical variety of different atomic scale junctions' geometries. Differential conductance measurements (dI/dV versus V) are conducted using a standard lock-in technique. A reference sine signal of 1 mV peak-to-peak voltage (Vpp) at about 3 kHz modulating a d.c. bias voltage is generated by the DAQ card. The alternating-current (a.c.) response is recorded by the DAQ card and extracted by a LabView implemented lock-in analysis to obtain the differential conductance as a function of bias voltage.

Extended Data Fig. 2 shows the electronic setup connected to the sample. The circuit can be switched between a conductance mode, which is used to measure the d.c. conductance of the examined junction and the dI/dV spectra, and a noise mode, applied to measure the noise generated by the junction. In the latter mode of measurement, the relatively noisy instruments used in the conductance mode are disconnected from the sample owing to the high sensitivity of the noise measurements. The voltage noise is amplified by a custom-made differential low-noise amplifier. The amplifier was calibrated by the thermal noise that is generated in a set of well-characterized resistors embedded in liquid nitrogen. A power spectrum between 0.25 kHz and 300 kHz is measured via a NI-PXI5922 DAQ card using a LabView implemented fast Fourier transform analysis and averaged 1,000 times. To assess the stability of our noise amplifier, we recorded the thermal noise temperature of junctions with different conductance values at the base temperature of the system in intervals of about 7 h. We did not observe any detectable shift in the obtained temperature. To measure shot noise, the sample is current-biased by a Yokogawa GS200 SC voltage source connected to the sample through two 1 MΩ resistors located in proximity to the sample. The total cabling length was minimized to reduce stray capacitance to about 40 pF. The low level of the measured noise signal from the sample makes it sensitive to extrinsic noise. To impede noise pickup, the measurement setup is located within a Faraday cage and all instruments are connected to a specially assigned quiet ground, and are optically isolated from a control computer outside the Faraday cage. All amplifiers are powered by batteries to avoid noise injection from power lines. Additionally, an RC filter (where R is resistance and C is capacitance) is connected after the piezo driver to minimize possible excitation of mechanical noise coupled to the junction through the piezoelectric element. The RC filter is bypassed when recording conductance traces in order to avoid interference with the measurements.

The temperature of each electrode is controlled by a feedback loop consisting a custom-made proportional–integral–derivative (PID) controller that is powered by batteries and located inside the Faraday cage, a heating resistor (thin-film, 100 Ω, Panasonic) that is thermally connected to each Au electrode by a sapphire housing, as well as a thermometer (Lakeshore DT-670 calibrated silicon diode) located at each electrode's tip (Fig. 1a). This feedback circuit is optically isolated from the control computer outside the Faraday cage. The system reaches an optimal stabilization around the preset temperature in about 35 min. Owing to the constant operation of a feedback loop, the actual temperature oscillates by at most 0.025 K around the set value. These variations are taken into consideration in the error calculations.

Molecular junction characterization. The typical conductance of gold atomic junctions is around $1G_0$, carried by one dominant transmission channel[28,29]. Since gold junctions with conductance below $0.7G_0$ frequently cannot be stabilized, we introduced hydrogen to form stable molecular junctions with a wider conductance range[23,24] below $1G_0$. This conductance window is necessary for the demonstration of the delta-T noise in Fig. 2. Before the introduction of molecules, the bare Au junction is characterized by constructing conductance histograms, as presented in Extended Data Fig. 1 (brown). The peak at $1G_0$, and the tail at low conductance are regarded as the typical fingerprints of a bare Au atomic junction[27,28]. The peak indicates the most probable conductance of a single atom Au junction (a single Au atom in the cross-section of the junction's constriction), while the tail at low conductance is the outcome of tunnelling conductance, measured after the breaking of a single atom junction. The blue conductance histogram exemplifies the different characteristics that emerge following the introduction of hydrogen. The large number of counts below $1G_0$ indicate the repeated formation of different stable molecular junction configurations with a broad range of conductance values. This feature of the studied molecular junctions allows us to perform noise measurements on stable junction configurations with a broad range of conductance below $1G_0$ (see Fig. 2 and Extended Data Fig. 8 for noise data obtained for hydrogen based molecular junctions and bare atomic gold junctions, respectively). In our setup, shot noise and delta-T noise give less reliable results below $0.1G_0$ owing to RC low-pass filtering. Therefore, we limit our analysis to junctions with conductance above $0.1G_0$.

Calibration of thermometers by thermal noise. Thermal noise identifies the electronic temperature that determines the Fermi–Dirac distribution of electrons in the electrodes (usually in a region of tens to hundreds of nanometres around the atomic scale junction). The silicon diode thermometers are attached to the surface of the electrodes near the electrode tips (Fig. 1a). The electric wires of these thermometers are anchored to metal thermalization plates at about 4.2 K to prevent the absorption of heat from their hot side, which is located outside the cryostat. As a result, when the sample is heated above the base temperature, the detected temperature by the thermometers is always lower than the thermal noise temperature. With the aid of thermal noise measurements, we could calibrate the temperature indicated by the thermometer to give the actual temperature in the nanoscale vicinity of the studied junction. The thermal noise was measured as a function of conductance at several fixed temperatures. The inset of Extended Data Fig. 3 provides an example for such a measurement at 37.10 ± 0.04 K (determined by a linear fit to the thermal noise). Then, the relation between the temperature given by the thermometers and the temperatures given by thermal noise was extracted for the relevant temperature range in our experiment (Extended Data Fig. 3). We note that the difference between the temperature measured by the thermometers and the temperature extracted from the thermal noise is nullified at the base temperature of the setup, as observed in Extended Data Fig. 3. This is due to the fact that at the base temperature, the thermalization plates that are connected to the wires of the thermometers have the same temperature as the junction, and they do not cool the thermometers to a lower temperature with respect to the junction's temperature. The calibration procedure described was used to relate a temperature at the nanoscale vicinity of the junction to the thermometer reads. This calibration reliably evaluates the electronic temperature at the electrode apexes, using macroscale thermometers. This procedure was performed before each experiment.

Shot noise measurements and shot noise analysis of Au/H_2 junctions. Shot noise measurements on Au/H_2 junctions were performed as described in refs [20,29]. Extended Data Fig. 4 presents the Fano factor, which is defined in the main text but is also equal to the measured shot noise in units of $2eI$ (where I is the current). The Fano factor is presented as a function of the corresponding conductance; both are obtained for different realizations of the Au/H_2 junctions. Junctions that are characterized by conductance above $1G_0$, which is the typical conductance of gold single-atom junctions, are obtained by squeezing the two electrodes against each other to form contacts with more than one Au atom at the narrowest cross-section of the contact, possibly contaminated by hydrogen.

The red curve in Extended Data Fig. 4 indicates the minimal Fano factor that is obtained by a sequential opening of channels, which can be described as follows. The first channel gradually opens between 0 and $1G_0$, with a transmission probability being equal to the conductance. The second channel opens above $1G_0$, with a transmission probability of $G/G_0 - 1$, since the first channel is kept fully open

($\tau_1 = 1$), and so on. Therefore, molecular junctions with a Fano factor versus conductance that are located on the red curve below $1G_0$ are characterized by a single transmission channel. In fact, the bulk of the data below $1G_0$ accumulates close to this curve, indicating that the electronic transport of most molecular junctions is dictated by a single transmission channel with minor contributions from secondary channels (see inset I in Extended Data Fig. 4). Some of the data points at $1G_0$ show full suppression of the Fano factor. In these cases, the conductance is determined by a single channel with transmission probability of one. The dashed black line, which practically serves as the upper limit for the measured data below $1G_0$, indicates the maximal Fano factor that can be obtained for two channels (along the dashed line, the two channels have equal transmission probability). The Fano factor is insensitive to the conductance above $4G_0$ (ref. [29]), which is the relevant range for the thermal noise analysis presented in Fig. 1. In this conductance range, the Fano factor scatters around the averaged value of 0.28 with a standard deviation of 0.07.

Ratio between delta-T noise and thermal noise above $4G_0$. For $G > 4G_0$ we can assume a constant Fano factor ($F = 0.2$–0.4). The ratio between the delta-T noise ($S_{\Delta T}$) and the thermal noise (S_{TN}) is

$$\frac{S_{\Delta T}}{S_{TN}} = \frac{k_B G_0 F[(\Delta T)^2/\bar{T}](\pi^2/9 - 2/3)}{4k_B G_0 \bar{T}} = \frac{F}{4}\left(\frac{\Delta T}{\bar{T}}\right)^2\left(\frac{\pi^2}{9} - \frac{2}{3}\right) \tag{3}$$

where

$$\begin{aligned} \left.\frac{S_{\Delta T}}{S_{TN}}\right|_{F=0.2} &= 0.0215\left(\frac{\Delta T}{\bar{T}}\right)^2 \\ \left.\frac{S_{\Delta T}}{S_{TN}}\right|_{F=0.4} &= 0.043\left(\frac{\Delta T}{\bar{T}}\right)^2 \end{aligned} \tag{4}$$

Thus, a maximal ratio of 5% is obtained in our measurements for the extreme case of $\Delta T = 22.9$ K, $\bar{T} = 21.3$ K, and $F = 0.4$. However, typically this ratio is about 3%.

Measurements of noise at finite temperature difference. Following the formation of an atomic scale junction with a fixed inter-electrode distance at a given temperature difference, a current versus voltage curve (Extended Data Fig. 5a) was measured and the conductance was determined from the curve's slope ($G = I/V$) at its linear regime around zero voltage. The noise at a given temperature difference was measured by switching to the noise circuit (Extended Data Fig. 2), and measuring the total noise versus frequency (Extended Data Fig. 5b). To ensure the stability of the junction during this process, a second current versus voltage measurement was performed right after the noise measurement by switching back to the conductance circuit. The entire procedure of the two current versus voltage measurements and noise measurement takes about 30 s. Only when the difference between the conductance values found before and after the noise measurement was less than about 1% was the noise measurement considered valid. The voltage noise of the measurement setup was measured separately for a fully formed junction (a short circuit) at the same ΔT and was subtracted from the total noise spectra obtained in the experiment. The typical voltage noise of our setup varies between 8.1×10^{-19} and 9.0×10^{-19} V^2 Hz^{-1} (0.90–0.95 nV $Hz^{-1/2}$). Extended Data Fig. 5b presents the measured total noise for a set of junctions that are characterized by different conductance values, given by the different slopes of the curves presented in Extended Data Fig. 5a.

The suppression of the noise as a function of frequency observed in Extended Data Fig. 5b is the outcome of low-pass RC suppression due to the finite resistance (R) and capacitance (C) of our setup. Furthermore, this noise contains a small yet finite contribution from the amplifier input current noise (S_I^{in}) that is also subject to RC suppression. To account for these two effects, C and S_I^{in} were determined by optimally fitting an RC function (taking into account the current noise contribution) to thousands of noise versus frequency spectra measured at different conductance (for example, Extended Data Fig. 5c) and temperature (for example, Extended Data Fig. 5d) in the relevant range of our analysis ($0.1G_0$–$7.0G_0$ and 5.4–50.4 K). The noise spectra were fitted to the following RC transfer function S (in units of V^2 Hz^{-1}):

$$S = \frac{S_0}{1 + (2\pi f R C)^2} \tag{5}$$

where f is the frequency and S_0 is the zero frequency total noise. The amplifier input current noise was obtained (in units of V^2 Hz^{-1}) by

$$S_0 = 4k_B T R + [S_I^{in}(f)]^2 R^2 \tag{6}$$

The term $4k_B TR$ is the voltage thermal noise (note that $\Delta T = 0$ during this procedure). The typical capacitance of our measurement system is $C = 42.4 \pm 0.1$ pF and the amplifier input current noise (in units of A^2 Hz^{-1}) is $S_I^{in}(f) = 1.37 \times 10^{-32} \times f$, which has a linear dependence on frequency.

Once C and S_I^{in} are determined, every total noise spectrum that is measured at a finite temperature difference (Extended Data Fig. 5b) is corrected by the inverse of the RC function, using the obtained resistance from conductance measurements ($R = 1/G$; Extended Data Fig. 5a). S_I^{in} is then subtracted from the total noise to obtain the corrected total noise, presented in Extended Data Fig. 5e). Finally, every noise spectrum is averaged in a selected frequency window of 180–230 kHz, as seen in Extended Data Fig. 5e (the results are not sensitive to the selected range). The average values of the total noise as a function of conductance appear in Extended Data Fig. 5f, where the units are converted to A^2 Hz^{-1} by dividing each averaged value by the square of the corresponding resistance.

The contribution of shot noise generated by thermovoltage. To reveal the contribution of shot noise due to the generated thermovoltage in our measurement setup, the total thermovoltage of the system (sample and wires) was measured at the maximal temperature difference considered in Figs. 2 and 3. The measurement procedure is based on the technique described in ref. [30]. In Extended Data Fig. 6a we present the measured total thermovoltage as a function of conductance for the Au/H_2 junctions at $\Delta T = 25.3 \pm 0.6$ K and $\bar{T} = 26.3 \pm 0.7$ K. The scattering of the total thermovoltage for different junctions is more pronounced below $1G_0$, probably owing to the increased sensitivity of the transmission dependence on energy to structural variations when an atomic constriction is formed in the junction. A similar increase in the scattering of the data below $1G_0$ was observed in thermopower measurements on bare gold atomic junctions[31].

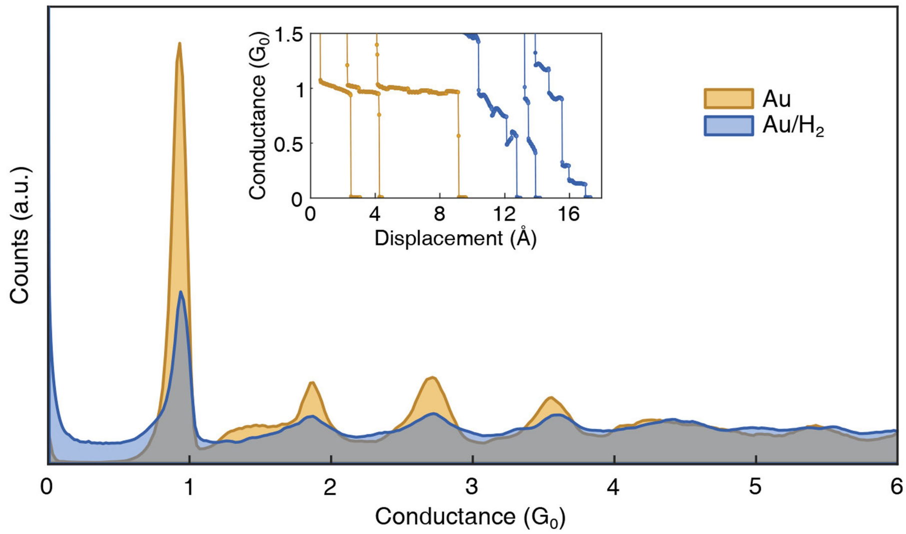
To estimate the maximal shot noise that can be generated in our experiments, we used equation (1) to calculate the shot noise that is expected for 155 μV, which is the largest thermovoltage acquired in our measurements (red star in Extended Data Fig. 6a). The obtained shot noise is plotted in red in Extended Data Fig. 6b along with the measured excess noise at the same average temperature and temperature difference. The largest expected shot noise due to the generation of thermovoltage in the junction is about three orders of magnitude smaller than the measured excess noise at a finite temperature difference. Thus, the observed excess noise in our experiments is not an outcome of the standard shot noise, and the fraction of shot noise contribution to this noise is practically negligible.

Fano factor based on delta-T noise measurements. Since the delta-T noise (second term in equation (2)) depends on $\sum_i \tau_i(1-\tau_i)$, the Fano factor $F = \sum_i \tau_i(1-\tau_i)/\sum_i \tau_i$ can be obtained from conductance ($G = G_0\sum_i \tau_i$) and delta-T noise measurements, rather than by measuring the standard voltage-activated shot noise, which is the usual approach. Extended Data Fig. 9 presents the Fano factor versus conductance, obtained from noise measurements at a finite temperature difference and zero applied voltage. The overall behaviour is similar to the one presented in Extended Data Fig. 4. This agreement serves as an additional indication that the excess noise is in fact due to the delta-T noise. Furthermore, it illustrates that information about the distribution of transmission channels can be obtained from delta-T noise measurements.

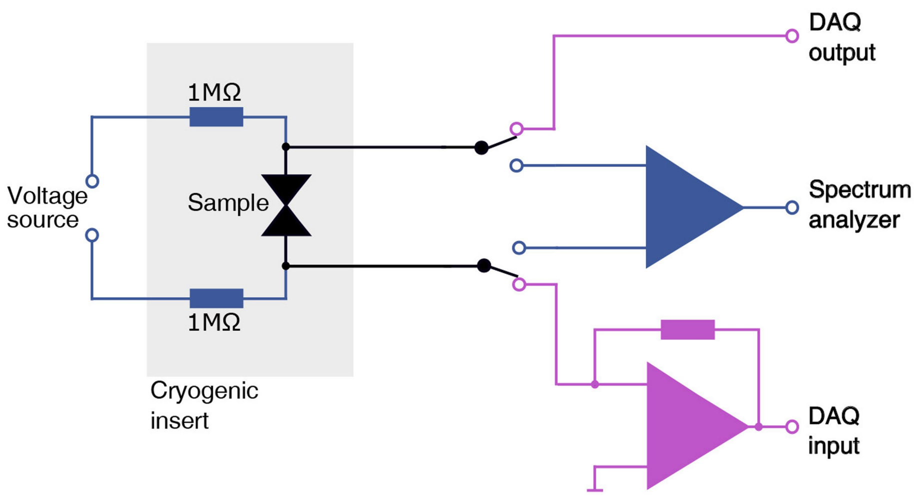
Data availability

The datasets generated and analysed during this study are available from the corresponding author on reasonable request.

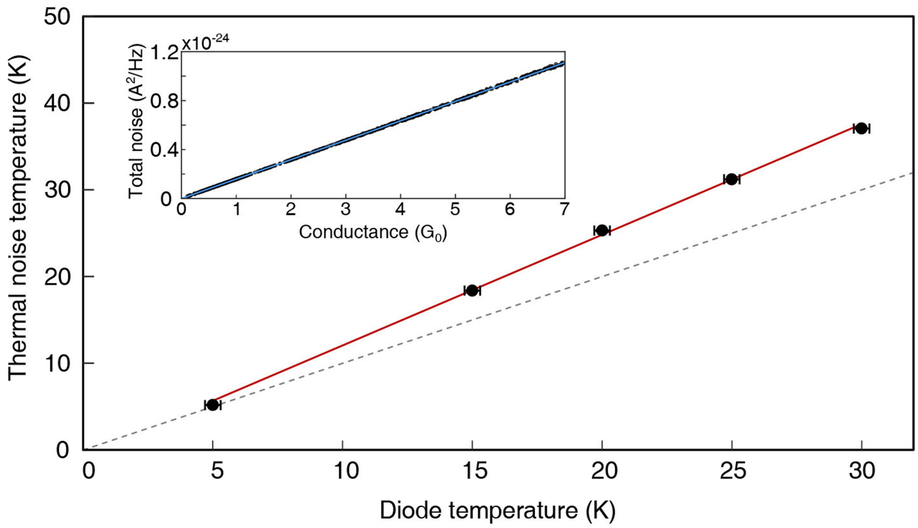
27. Ludoph, B., Devoret, M. H., Esteve, D., Urbina, C. & Van Ruitenbeek, J. M. Evidence for saturation of channel transmission from conductance fluctuations in atomic-size point contacts. *Phys. Rev. Lett.* **82**, 1530 (1999).
28. Rubio-Bollinger, G. et al. Single-channel transmission in gold one-atom contacts and chains. *Phys. Rev. B* **67**, 121407 (2003).
29. Vardimon, R., Matt, M., Nielaba, P., Cuevas, J. C. & Tal, O. Orbital origin of the electrical conduction in ferromagnetic atomic-size contacts: insights from shot noise measurements and theoretical simulations. *Phys. Rev. B* **93**, 085439 (2016).
30. Evangeli, C. et al. Engineering the thermopower of C_{60} molecular junctions. *Nano Lett.* **13**, 2141–2145 (2013).
31. Ludoph, B. & Van Ruitenbeek, J. M. Thermopower of atomic-size metallic contacts. *Phys. Rev. B* **59**, 12290 (1999).
32. Vardimon, R., Klionsky, M. & Tal, O. Experimental determination of conduction channels in atomic-scale conductors based on shot noise measurements. *Phys. Rev. B* **88**, 161404 (2013).



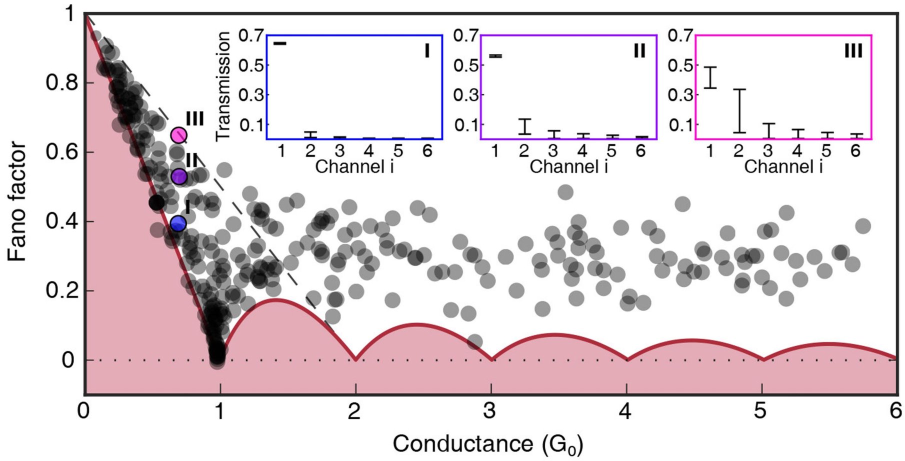
Extended Data Fig. 1 | Characterization of Au/H_2 molecular junctions. Conductance histograms of bare Au atomic junctions (brown) and Au/H_2 molecular junctions (blue) are shown. The histograms are composed from at least 1,500 conductance versus electrode displacement traces recorded at a bias voltage of 100 mV. a.u., arbitrary units.



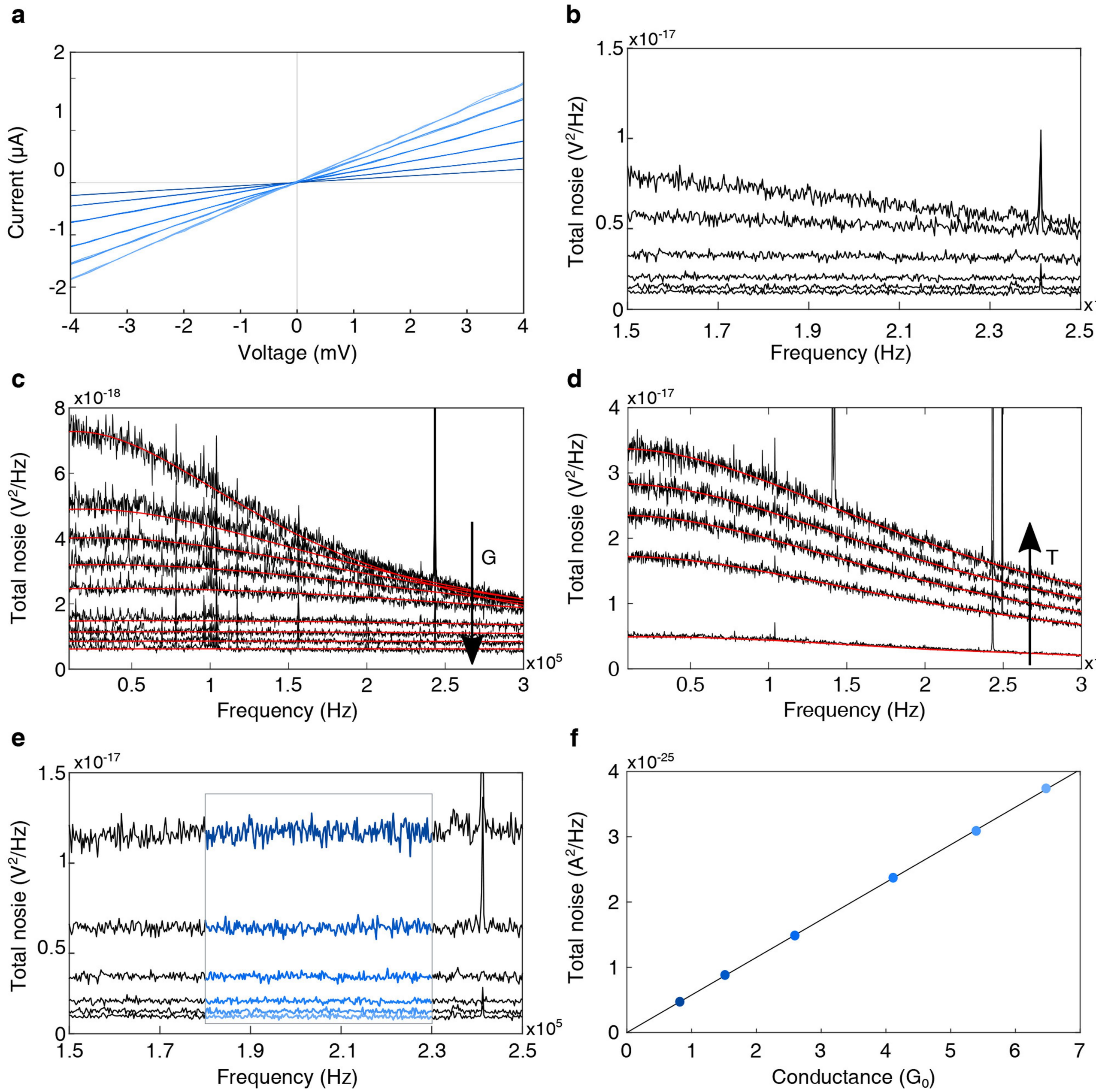
Extended Data Fig. 2 | Electronic measurement setup. Schematic presentation of the electronic circuit for conductance and noise measurements is shown. The electronic circuit consists of two switchable measurement circuits: a conductance circuit (purple) and a noise circuit (blue).



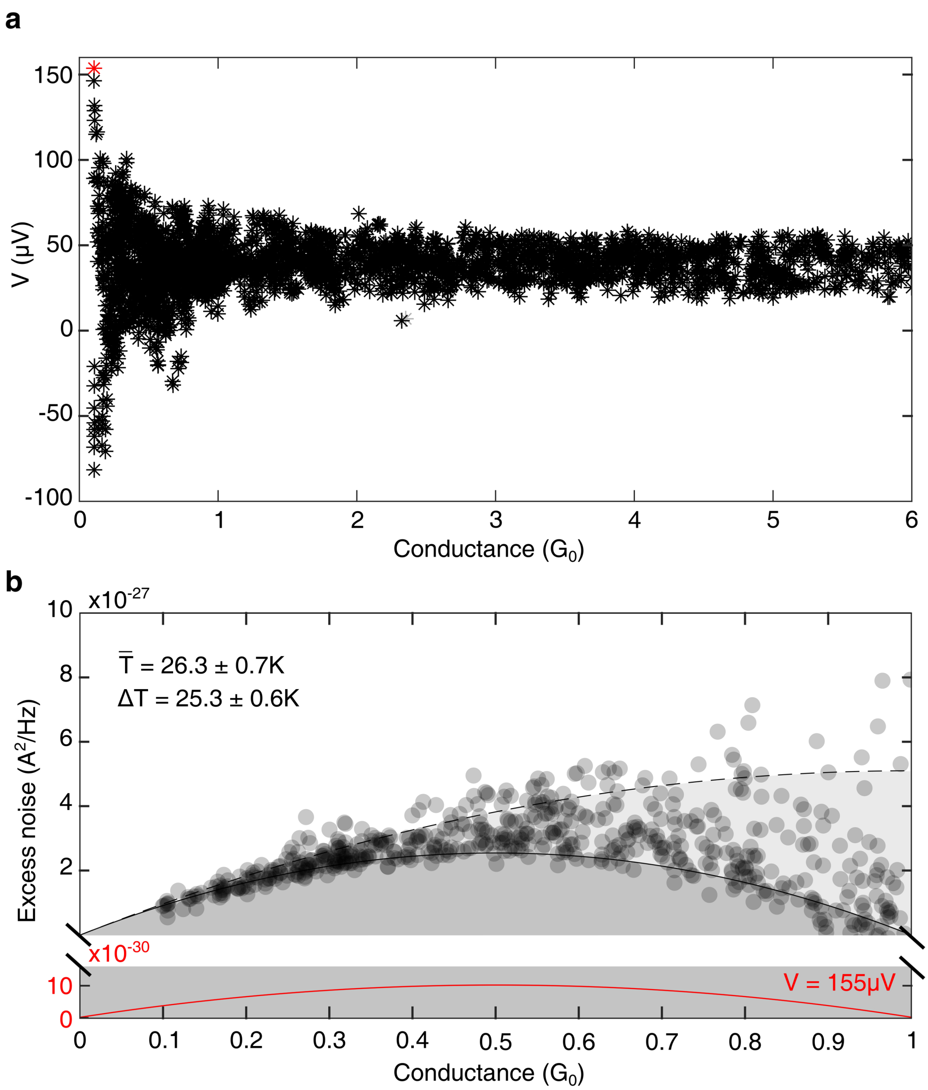
Extended Data Fig. 3 | Thermometer calibration based on thermal noise. The temperature measured by thermal noise is shown versus that measured using the diode thermometer (black circles; the vertical error bars are smaller than the circles' diameter). The error bars correspond to the systematic errors in our measurements. To guide the eye, the dashed grey line corresponds to a ratio of 1:1. The red line is a linear fit of the data. The calibration of the thermometers temperature is done by this fit $T_{\mathrm{TN}} = (1.28 \pm 0.02)T_{\mathrm{therm}} - 1.0 \pm 0.5\,\mathrm{K}$, where T_{therm} is the temperature measured by the thermometer. The inset shows an example for measured thermal noise versus conductance (black dots) at a thermal noise temperature of 37.10 ± 0.04 K. The blue line is a linear fit from which the thermal noise temperature is determined. This measurement procedure is repeated at different temperatures to construct the main graph. When the junction is heated above the setup base temperature, the thermometers attached to the electrode tips always indicate lower temperatures than those determined by the thermal noise.



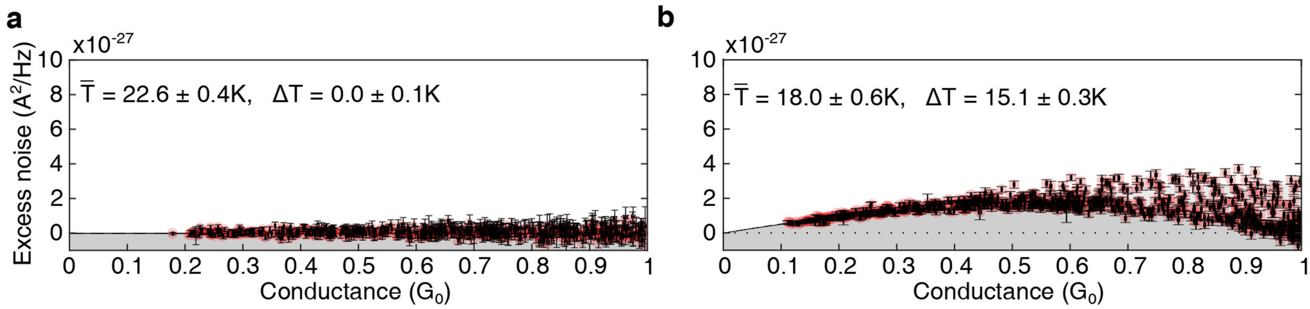
Extended Data Fig. 4 | Shot noise analysis for Au/H_2 junctions. The Fano factor extracted from shot noise and conductance measurements[20,29] is shown versus the conductance for different junction realizations at 4.6 K ($\Delta T = 0$). The thick red curve provides the minimal Fano factor. Data that accumulate on this line below $1G_0$ indicate junctions with a single transmission channel[22]. The dashed line provides the maximal Fano factor that two channels can generate for the relevant conductance. The insets show transmission probabilities of the main six transmission channels based on numerical analysis of the measured Fano factor and conductance[32] for the three marked cases (I, II, III) in the main panel. The error bars provide the range of transmission solutions that satisfies the measured conductance and shot noise. Inset I shows that a junction that is characterized by Fano factor and conductance data near the red curve conducts via a single dominant channel with only minor contribution from a secondary channel. In contrast, inset III exemplifies that a junction with Fano factor and conductance data near the dashed curve can conduct via two dominant channels with possible minor contributions from other channels.



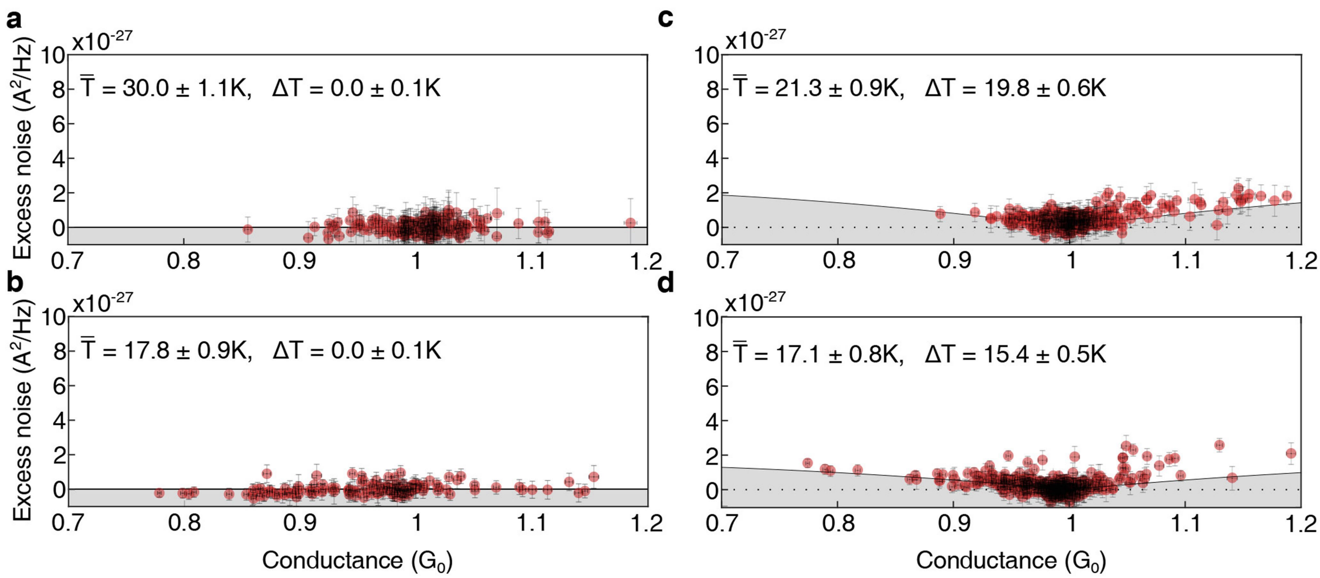
Extended Data Fig. 5 | Noise measurements at finite temperature differences. **a**, Current–voltage curves for a set of different junction realizations at $\bar{T} = 13.3 \pm 0.3$ K and $\Delta T = 12.6 \pm 0.2$ K. The conductance of each junction is obtained by the slope of the curve. Here, $G = 0.82G_0$, $1.52G_0$, $2.57G_0$, $4.05G_0$, $5.30G_0$ and $6.34G_0$, with all values $\pm 0.01G_0$, starting from the smallest slope. **b**, Total noise as a function of frequency for the same junctions examined in **a**. The top spectrum corresponds to the junction with the highest conductance. The noise is suppressed by low-pass *RC* filtering owing to the capacitance of the setup and the finite sample and wire resistance. **c**, Examples for *RC* transfer function and S_I^{in} fitting to spectra of total noise versus frequency measured at a fixed temperature of 5.4 ± 0.5 K, and different conductance values ($0.51G_0$–$6.03G_0 \pm 0.01G_0$). The arrow points in the direction of increasing conductance G. **d**, Same as **c** at a fixed conductance of $G = 0.77G_0 \pm 0.01G_0$, and different temperatures (5.4 ± 0.5 K to 37.5 ± 0.9 K). The arrow points in the direction of increasing temperature T. The setup capacitance and S_I^{in} are extracted from the fitting. **e**, The data presented in **b** corrected by an *RC* transfer function followed by subtraction of S_I^{in}. **f**, Total noise as a function of conductance obtained by averaging the noise presented in **e** in a frequency range of 180–230 kHz, coloured blue in **e**.



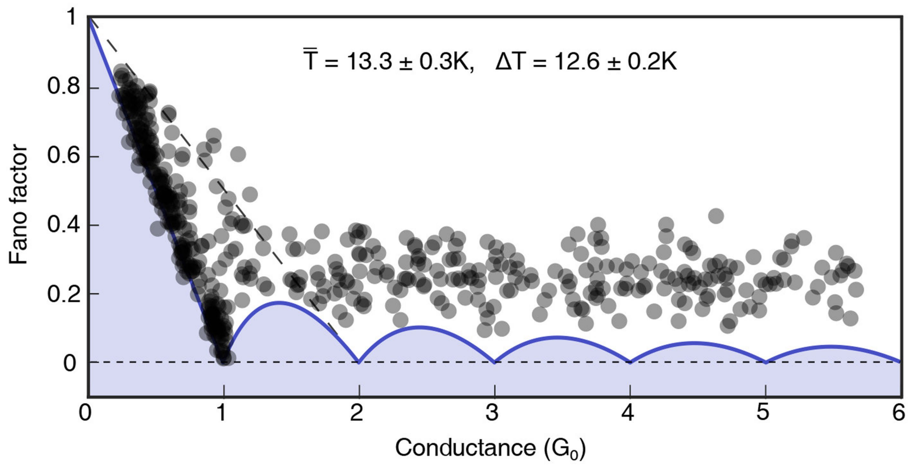
Extended Data Fig. 6 | Total thermoelectric voltage and its estimated contribution to the excess noise. a, Measured total thermoelectric voltage for the examined junctions as a function of conductance at $\bar{T} = 26.3 \pm 0.7$ K and $\Delta T = 25.3 \pm 0.6$ K. Measurements were performed according to the method described in ref. [30]. Here, we are interested in the total thermoelectric voltage that is built across the junction, since it can be a source for shot noise. **b**, Calculated shot noise (red curve) that is expected for the maximal measured thermovoltage in **a** (155 μV, marked as a red star in **a**), and the measured excess noise (dark circles) at the same $\bar{T}$ and ΔT for which the thermoelectric voltage was measured. Calculated delta-T noise is given by the black curve for a single transmission channel, and by the dashed curve for two channels with equal transmission probabilities (non-approximated numerical calculations based on equation (S2) in Supplementary Information).



Extended Data Fig. 7 | Excess noise measured at zero and finite temperature difference with error bars. a, **b**, Excess noise as a function of conductance measured in the examined molecular junctions as presented in Fig. 2a, b, including error bars, corresponding to the systematic errors in our measurements. The size of the error bars is comparable or slightly larger than the diameter of the semitransparent red symbols.



Extended Data Fig. 8 | Excess noise measured at zero and finite temperature difference for bare gold atomic junctions. a, **b**, Excess noise (obtained by subtracting the average thermal noise from the total measured noise) as a function of conductance measured in bare gold atomic junctions at different temperatures at thermal equilibrium ($\Delta T = 0$). **c**, **d**, Excess noise as a function of conductance measured at different average temperatures and finite temperature differences across the junctions ($\Delta T \neq 0$). Calculated delta-T noise is given by the black curve for single transmission channel probabilities (non-approximated numerical calculations based on equation (S2) in Supplementary Information). When a temperature difference is applied across the junctions, some enhancement of the excess noise is observed. The measured excess noise can be described by the theoretical expression for the delta-T noise, although the agreement is less clear than for hydrogen-based molecular junctions (Fig. 2), owing to the lack of data below $0.75G_0$. The spread in the results is a natural outcome of additional transmission channels that open as the conductance increases. The error bars correspond to the systematic errors in our measurements.



Extended Data Fig. 9 | Fano factor obtained from noise measurements at a finite temperature difference. The Fano factor (semitransparent black symbols) is extracted from the excess noise data presented in Fig. 2d, and the associated measured conductance using equation (2). The short-dashed horizontal line marks the zero Fano factor as a baseline. The thick blue curve provides the theoretically predicted minimal Fano factor. Data that accumulate on this line below $1G_0$ indicate junctions with a single transmission channel[22]. The long-dashed sloped line marks the maximal Fano factor that two channels can generate for the relevant conductance.

https://doi.org/10.1038/s41586-018-0575-3

Perovskite light-emitting diodes with external quantum efficiency exceeding 20 per cent

Kebin Lin[1], Jun Xing[2], Li Na Quan[3], F. Pelayo García de Arquer[3], Xiwen Gong[3], Jianxun Lu[1], Liqiang Xie[1], Weijie Zhao[2], Di Zhang[1], Chuanzhong Yan[1], Wenqiang Li[1], Xinyi Liu[1], Yan Lu[1], Jeffrey Kirman[3], Edward H. Sargent[3]*, Qihua Xiong[2]* & Zhanhua Wei[1]*

Metal halide perovskite materials are an emerging class of solution-processable semiconductors with considerable potential for use in optoelectronic devices[1–3]. For example, light-emitting diodes (LEDs) based on these materials could see application in flat-panel displays and solid-state lighting, owing to their potential to be made at low cost via facile solution processing, and could provide tunable colours and narrow emission line widths at high photoluminescence quantum yields[4–8]. However, the highest reported external quantum efficiencies of green- and red-light-emitting perovskite LEDs are around 14 per cent[7,9] and 12 per cent[8], respectively—still well behind the performance of organic LEDs[10–12] and inorganic quantum dot LEDs[13]. Here we describe visible-light-emitting perovskite LEDs that surpass the quantum efficiency milestone of 20 per cent. This achievement stems from a new strategy for managing the compositional distribution in the device—an approach that simultaneously provides high luminescence and balanced charge injection. Specifically, we mixed a presynthesized $CsPbBr_3$ perovskite with a MABr additive (where MA is CH_3NH_3), the differing solubilities of which yield sequential crystallization into a $CsPbBr_3$/MABr quasi-core/shell structure. The MABr shell passivates the nonradiative defects that would otherwise be present in $CsPbBr_3$ crystals, boosting the photoluminescence quantum efficiency, while the MABr capping layer enables balanced charge injection. The resulting 20.3 per cent external quantum efficiency represents a substantial step towards the practical application of perovskite LEDs in lighting and display.

$MAPbI_{3-x}Cl_x$ and $MAPbBr_3$ were used in early perovskite LEDs that achieved external quantum efficiencies (EQEs) of 0.76% and 0.1% for the near-infrared and green regimes, respectively[14]. Two strategies[5–7,15] have since led to notable improvements in LED performance. The first strategy involves direct spin-coating of colloidal perovskite nanocrystals[16–19]; these nanocrystals are highly luminescent with a photoluminescence quantum yield (PLQY) of nearly 90%, and their optical properties can be tuned by compositional engineering and crystal size. The second approach requires deposition of bulk perovskite films using perovskite precursor solutions whose composition can be suitably engineered (for example, stoichiometrically modified $MAPbBr_3$[6] and $MAPbX_3$ with the addition of long-chain ammonium halides[5] or 1-naphthylmethylamine halides[8]).

Here we have built on prior works and pursued a new strategy for generating still higher EQEs. Our approach was to combine a high PLQY with balanced charge injection by constructing a compositionally graded perovskite based on a quasi-core/shell structure: the bottom part consists of perovskite light-emitting polycrystals capped with a defect-passivation layer that passivates the grain boundaries; and the top part serves to passivate the surface and simultaneously balance charge injection into the perovskite LED device.

We used the very different solubility and crystallinity of perovskite ($CsPbBr_3$) and passivant (MABr) in polar solvent and—through a one-step deposition method—fabricated in situ the compositionally graded material. This consisted of a defect-passivated perovskite layer on the bottom ($CsPbBr_3$/MABr) and an electrical passivating layer (MABr) on top. The resulting highly luminescent perovskite films and balanced charge injection enabled the development of perovskite LED devices with an initial EQE of 17%. We further improved the device charge injection balance by inserting an insulating layer of poly(methyl methacrylate) (PMMA) between the perovskite layer and the electron-transfer layer (ETL), thereby maximizing the device efficiency at 20.3%.

We synthesized a $CsPbBr_3$ perovskite powder as a starting material, and then added MABr. We engineered the amount of MABr additive to improve perovskite film formation and PLQY (for example, 'mixture 1.0' signifies that the molar ratio of MABr to $CsPbBr_3$ is 1). Figure 1a shows three different perovskite structures fabricated using a strategy that we term compositional distribution management: single-layered $CsPbBr_3$ (prepared by one-step spin-coating); bilayered $CsPbBr_3$/MABr (prepared by coating another layer of MABr on the as-formed $CsPbBr_3$ layer); and quasi-core/shell $CsPbBr_3$/MABr (mixture 1.0). Figure 1b and Extended Data Fig. 1 show that the capping layer of MABr (in the bilayered structure) only slightly enhances photoluminescence emission, while the mixture-1.0 film with quasi-core/shell structure presents very bright photoluminescence emission (see Supplementary Information, video S1). We found an enhancement of photoluminescence proportional to the amount of MABr additive when the molar ratio of MABr to $CsPbBr_3$ was increased from 0.4 to 1.0; photoluminescence began to decrease when the molar ratio was increased beyond 1.0 (Extended Data Fig. 1).

We observed that, during film formation, $CsPbBr_3$ crystallized rapidly, whereas MABr sequentially increased its crystallization rate after the $CsPbBr_3$ precursor was completely consumed (Extended Data Fig. 1d). We explain this by noting that the solubility of $CsPbBr_3$ (which we find to be 0.56 M) is far below than that of MABr (5 M) in dimethylsulfoxidde (DMSO).

Secondary ion mass spectrometry (SIMS) depth analysis (Fig. 1c) supports the $CsPbBr_3$/MABr gradient structure. From the SIMS results, one can see that the top layer contains CH_4N^+ ions (from the capping MABr; stage I); the middle layer comprises Pb^+ ions with few CH_4N^+ ions (from $CsPbBr_3$ and MABr; stage II); and the bottom layer consists of In^+ ions (from the indium tin oxide (ITO) used as the support; stage III). To gain insight into the compositional distribution of the as-formed mixture-1.0 films, we carried out cross-sectional scanning electron microscopy (SEM) and transmission electron microscopy (TEM) studies. We prepared the cross-sectional TEM samples by using a focused ion beam, with C and Pt layers predeposited in order to protect the perovskite from possible ion-beam-induced damage. The cross-sectional SEM image (Extended Data Fig. 2a) shows a high-quality perovskite film with obvious grain boundaries. The cross-sectional TEM and element-mapping images (Fig. 1d and Extended Data Fig. 2b) show a well defined layer-by-layer structure, with ITO at the bottom, topped with poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS), then with $CsPbBr_3$, MABr and finally C. A shell of

[1]Engineering Research Center of Environment-Friendly Functional Materials, Ministry of Education, College of Materials Science & Engineering, Huaqiao University, Xiamen, China. [2]Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Singapore. [3]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada. *e-mail: weizhanhua@hqu.edu.cn; qihua@ntu.edu.sg; ted.sargent@utoronto.ca

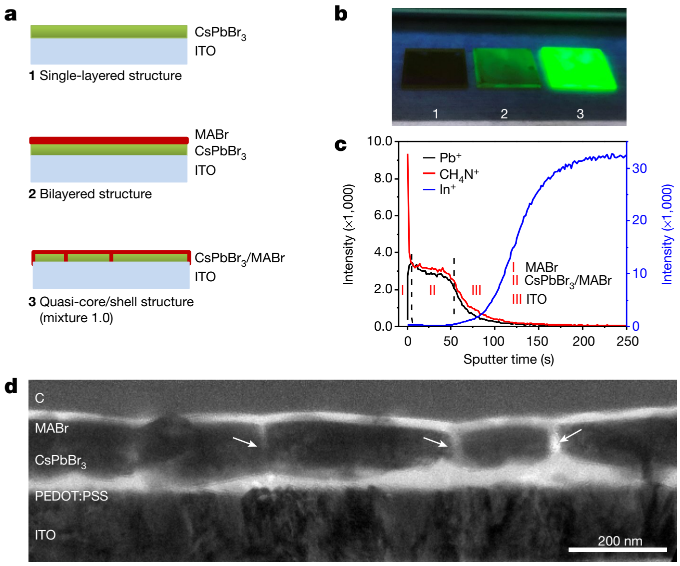


Fig. 1 | Enhancing photoluminescence through compositional distribution management. **a**, Schematic illustrations of single-layered $CsPbBr_3$, bilayered $CsPbBr_3$/MABr, and quasi-core/shell $CsPbBr_3$/MABr structures, all fabricated on ITO substrates. **b**, Photographs of the three as-prepared perovskite films under ultraviolet light. **c**, SIMS depth analysis of the as-prepared quasi-core/shell $CsPbBr_3$/MABr structure on ITO glass. **d**, Cross-sectional TEM image of the quasi-core/shell $CsPbBr_3$/MABr structure on PEDOT:PSS. White arrows indicate the MABr shell (the grain boundary). The sample was prepared using a focused ion beam, and the top C layer was predeposited in order to protect the perovskite.

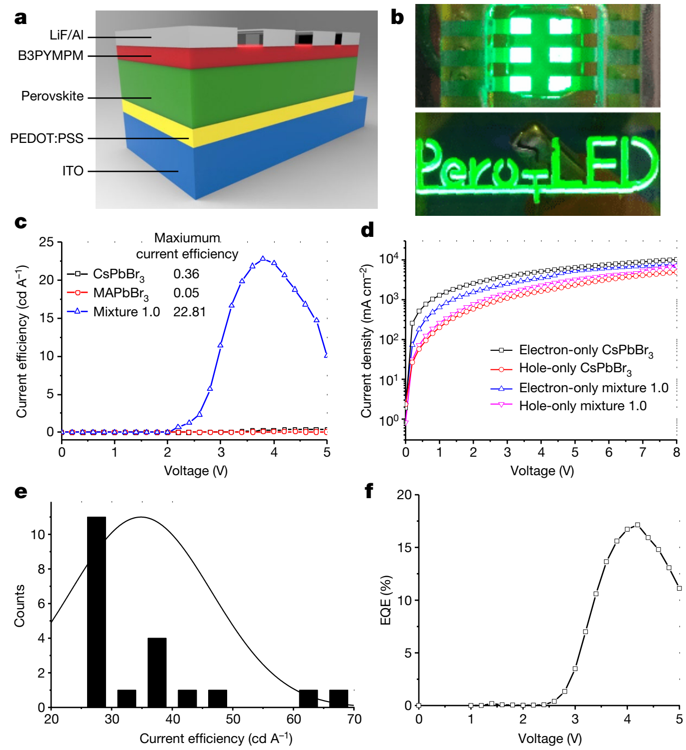


Fig. 2 | Fabrication of perovskite LEDs and performance evaluation. **a**, Configuration of a perovskite LED cell, with PEDOT:PSS and B3PYMPM as the HTL and ETL, respectively. **b**, Photographs of perovskite LED devices fabricated with the mixture-1.0 perovskite, showing six uniform and bright pixels and a logo of 'Pero-LED'. **c**, Typical current efficiency–voltage curves of the three perovskite LEDs, without optimization. **d**, Current density–voltage (J–V) curves of electron-only and hole-only devices. **e**, Current efficiency statistics for the mixture-1.0 perovskite LED. **f**, EQE–voltage characteristics of the best-performing mixture-1.0 perovskite LED.

MABr can be seen in the grain boundaries of $CsPbBr_3$ (white arrows in Fig. 1d), and another layer of MABr caps the $CsPbBr_3$, forming the quasi-core/shell structure. We sought to estimate experimentally the trap state density of the three classes of perovskite samples. We found (Extended Data Fig. 3a) that the MABr shell reduces defects in the mixture-1.0 perovskite films by a factor of four compared with the single-layered $CsPbBr_3$ film.

To gain insight into the effect of the thick upper layer of MABr on photoluminescence enhancement, we washed the bright perovskite film using anhydrous isopropyl alcohol (IPA) solvent, and observed that the photoluminescence decreased gradually as the MABr was removed (Extended Data Fig. 3b). To make a direct comparison, we also prepared pure $CsPbBr_3$ and $MAPbBr_3$ perovskite films. As shown in Extended Data Fig. 3c, all perovskite films show a transparent yellow colour under room illumination. However, only the mixture-1.0 perovskite films reveal high brightness under ultraviolet-lamp excitation. The ultraviolet/visible absorbance spectra (Extended Data Fig. 3d) of the mixture-1.0 film present a band-edge absorbance at 531 nm, similar to $CsPbBr_3$ (528 nm), corresponding to a bandgap of 2.33 eV. Analysis of the photoluminescence spectra of the three perovskite samples indicates that emission from the mixture-1.0 film is close to the emission of pure-$CsPbBr_3$ films (Extended Data Fig. 3e). In order to quantify the photoluminescence enhancement induced by the MABr additive, we measured the absolute PLQY according to a reported protocol[20]. We determined that the PLQY of the mixture-1.0 perovskite film was about 80%, while the PLQYs of $CsPbBr_3$ and $MAPbBr_3$ are not detectable (from an analysis of our system signal-to-noise ratio, we conclude that their PLQYs lie below 1%). Time-resolved photoluminescence spectra (Extended Data Fig. 3f) show that the mixture-1.0 film has a 50% longer radiative lifetime than that of pure $CsPbBr_3$, and that the longer lifetime of the photoluminescence transition is direct evidence of a decrease in the concentration of defects and an increase in film crystallinity[1,7,9]. We attribute the long lifetime of photoluminescence to the fact that the compositionally graded structure combined with the MABr shell passivates the nonradiative defects in $CsPbBr_3$.

We analysed the crystal structure of the perovskite films using X-ray diffraction (XRD; Extended Data Fig. 4a). We conclude that the mixture-1.0 perovskite exhibits the same crystal structure as monoclinic $CsPbBr_3$, instead of a mixture of separate phases of $CsPbBr_3$ and $MAPbBr_3$. X-ray photoelectron spectroscopy (XPS) data (Extended Data Fig. 4b, c) also indicate the existence of $CsPbBr_3$ and MABr in the mixture-1.0 perovskite film.

The perovskite layer requires high surface coverage and low roughness in order to achieve high-performance LEDs. We characterized the surface morphology of the pure $CsPbBr_3$, $MAPbBr_3$ and mixture-1.0 perovskite films using SEM and atomic force microscopy (AFM) (Extended Data Fig. 5). We observed small particles and pinholes in the $CsPbBr_3$ and $MAPbBr_3$ films; by contrast, in the mixture-1.0 film, smooth and well packed micrometre-sized cuboids were combined with good crystallinity.

We fabricated perovskite LEDs consisting of single-layered $CsPbBr_3$, single-layered $MAPbBr_3$, bilayered $CsPbBr_3$ and MABr, or mixture-1.0 perovskites with a quasi-core/shell structure, based on a device structure consisting of layered ITO/PEDOT:PSS/perovskite/B3PYMPM/LiF/Al (Fig. 2a, where B3PYMPM is $C_{37}H_{26}N_6$). PEDOT:PSS served as the hole-transfer layer (HTL), B3PYMPM as the ETL, LiF as an electron-injection layer and Al as the cathode (Extended Data Fig. 6a). Photographs of mixture-1.0 perovskite LED devices with six uniform and bright green-emitting pixels (2 mm × 1.5 mm) are shown in Fig. 2b. Larger-area devices (6 mm × 20 mm; Extended Data Fig. 6b) showcase uniform and bright emission.

The device performance of bilayered $CsPbBr_3$/MABr perovskite is quite limited (Extended Data Fig. 6c)—only slightly better than that of the single-layered $CsPbBr_3$—possibly because of its low PLQY and

poor surface morphology. The mixture-1.0 devices display an emission peak at 525 nm with a full width at half maximum (FWHM) of 20 nm (Extended Data Fig. 6d, e), corresponding to a CIE colour-space coordinates of (0.18, 0.75). We collected current-density–voltage (J–V) and luminance–voltage (L–V) curves in order to evaluate the LED performance (Extended Data Fig. 6f–h). Devices fabricated using mixture 1.0 exhibited the lowest current density at the same time as the highest luminance, indicating the best performance among our three classes of perovskite LED. The mixture-1.0 devices gave a maximum current efficiency of 23 cd A^{-1} at 3.8 V—fully three orders of magnitude higher than the current efficiency of the pure $CsPbBr_3$- and $MAPbBr_3$-based LEDs (Fig. 2c).

We posited that the superior LED performance of the mixture-1.0 devices arises not only from the high PLQY, but also from its combination with improved charge injection balance. To quantify charge injection, we measured the J–V scharacteristics of electron-only devices (ITO/B3PYMPM/perovskite/B3PYMPM/Al) or hole-only devices (ITO/PEDOT:PSS/perovskite/Au) (Fig. 2d). We conclude that electrons dominate injection into the pure-$CsPbBr_3$ devices, whereas a more balanced charge injection occurs in the case of mixture-1.0 devices. After optimization, we obtained an average current efficiency of 35 cd A^{-1} from 20 devices, with the best current efficiency reaching 65 cd A^{-1} (Fig. 2e). The EQE–V characteristics of the best-performing device show a maximum EQE of 17% at 4.2 V (Fig. 2f)—a record for a green-emitting perovskite LED.

Figure 2d shows that the capping MABr layer in the quasi-core/shell structure helps to reduce electron injection and improve charge balance. We thought that further improvement could potentially be realized through additional optimization of charge balance. We achieved this by depositing a thin PMMA layer on the as-formed perovskite. We then tested electron-only and hole-only device performance again, and found that the PMMA layer further helps in balancing charge injection (Fig. 3a). We therefore inserted a thin PMMA layer between the perovskite and ETL (Fig. 3b and Extended Data Fig. 7a). We found the PMMA to be continuous and smooth (Extended Data Fig. 7b, c), enabling charge injection into perovskite via tunnelling[13]. After we optimized the thickness of the PMMA layer and the molar ratio between MABr and $CsPbBr_3$ in the mixed perovskite precursor (Extended Data Fig. 7d, e), the devices reached a higher current efficiency of 78 cd A^{-1} (Fig. 3c).

Figure 3d presents the J–V and L–V curves of the best-performing mixture-1.0 device, showing a low driving current density and high luminance of 14,000 cd m^{-2}. A low turn-on voltage of 2.7 V—just slightly higher than the bandgap of the mixture-1.0 perovskite—is obtained because of the high quality of the perovskite thin film and the more efficient carrier injection from the HTL and ETL. We also found that the electroluminescence spectra at different applied voltages remained the same, and that the maximum power efficiency was 69 lm W^{-1} at 3.6 V (Extended Data Fig. 8). A maximum EQE value of 20.3% is achieved with a luminance of 3,400 cd m^{-2} (Fig. 3e and Supplementary Information, videos S2 and S3).

We measured the lifetime of the device by applying a constant current and monitoring the evolution of luminance[13,21]. After we applied a constant driving current of 5 mA (167 mA cm^{-2}), the luminance increased from 3,800 cd m^{-2} to 7,130 cd m^{-2} (L_0) in 0.66 min and then began to diminish (Fig. 3f). The half-lifetime (T_{50})—defined as the time taken for the luminance to decrease to $L_0/2$—was about 10 min. By using a calculation of $L_0^n T_{50} =$ constant, and assuming an acceleration factor of $n = 1.5$ (ref. [13]; Extended Data Fig. 9a), we estimate this device's T_{50} at 100 cd m^{-2} to be about 100 h—to our knowledge, the highest value estimated to date in high-performance perovskite LEDs, and an important step towards practical application. We also measured the stability of the device under continuous operation with luminance maintained at a constant value of about 100 cd m^{-2}, achieved by tuning the applied current to maintain luminance. The results (see Supplementary Information, video S4, and Extended Data Fig. 9b) show that the device could operate continuously for about 46 h, a similar order of magnitude to the extrapolated value from the accelerated ageing test. The corresponding EQE decreased from 13% to 5.6% over these 46 hours of continuous operation, indicating that the device did degrade even at this low luminance. Nonetheless, the stability performance shown here is two to three orders of magnitude higher than previous reported values[7,9,22–25] (Extended Data Table 1).

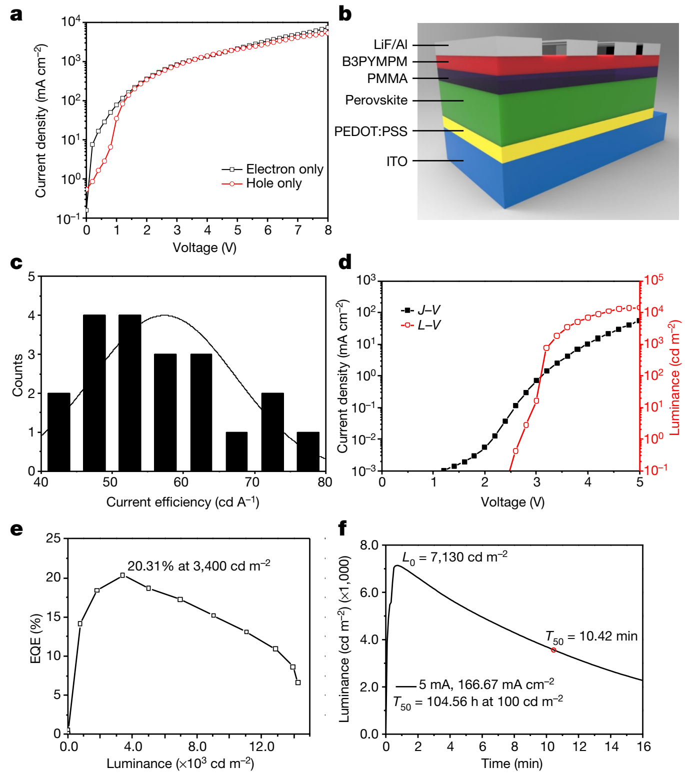


Fig. 3 | Enhancing the performance of perovskite LEDs by inserting a thin PMMA layer between the perovskite and the ETL. **a**, Current density–voltage (J–V) curves of electron-only and hole-only devices with a PMMA layer. **b**, Configuration of a perovskite LED cell with a very thin PMMA layer inserted between the perovskite and the ETL. **c**, Histogram showing current efficiency statistics of perovskite LED devices with a PMMA layer. **d**, **e**, L–V and J–V (**d**) and EQE–L (**e**) characteristics of the best-performing perovskite LEDs. **f**, Lifetime measurements of the best-performing mixture-1.0 perovskite LED device. A constant driving current of 5 mA (167 mA cm^{-2}) led to the luminance increasing from 3,800 cd m^{-2} to 7,130 cd m^{-2} (L_0) and then diminishing. We estimate this device's T_{50} at 100 cd m^{-2} to be about 104.56 h.

In summary, we have demonstrated a new strategy for realizing compositionally graded perovskite devices that simultaneously achieve high PLQY and balanced charge injection. Our approach exploits the differing solubilities of perovskite precursors to control the crystallization of $CsPbBr_3$/MABr gradient structure in a single step. The MABr shell passivates nonradiative defect sites in $CsPbBr_3$ crystals, and the MABr capping layer balances charge injection. These effects together allow us to achieve perovskite LEDs with a narrow green emission, exhibiting a record EQE that surpasses 20%. This high EQE is now on a par with those of more mature technologies such as organic LEDs. Improvements in device stability remain a challenge, and strides in that direction might be achieved by suppressing ion migration with additives or a blocking layer, fabricating a high-quality perovskite layer, and further optimizing the perovskite/ETL and perovskite/HTL interfaces[26–28].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0575-3.

Received: 22 January; Accepted: 22 August 2018;
Published online 10 October 2018.

1. Yang, W. S. et al. Iodide management in formamidinium-lead-halide-based perovskite layers for efficient solar cells. *Science* **356**, 1376–1379 (2017).
2. Ha, S.-T., Shen, C., Zhang, J. & Xiong, Q. Laser cooling of organic–inorganic lead halide perovskites. *Nat. Photon*. **10**, 115–121 (2016).
3. Ning, Z. et al. Quantum-dot-in-perovskite solids. *Nature* **523**, 324–328 (2015).
4. Yuan, M. et al. Perovskite energy funnels for efficient light-emitting diodes. *Nat. Nanotechnol*. **11**, 872–877 (2016).
5. Xiao, Z. G. et al. Efficient perovskite light-emitting diodes featuring nanometre-sized crystallites. *Nat. Photon*. **11**, 108–115 (2017).
6. Cho, H. et al. Overcoming the electroluminescence efficiency limitations of perovskite light-emitting diodes. *Science* **350**, 1222–1225 (2015).
7. Zhang, L. et al. Ultra-bright and highly efficient inorganic based perovskite light-emitting diodes. *Nat. Commun*. **8**, 15640 (2017).
8. Wang, N. N. et al. Perovskite light-emitting diodes based on solution-processed self-organized multiple quantum wells. *Nat. Photon*. **10**, 699–704 (2016).
9. Yang, X. et al. Efficient green light-emitting diodes based on quasi-two-dimensional composition and phase engineered perovskite with surface passivation. *Nat. Commun*. **9**, 570 (2018); correction **9**, 1169 (2018).
10. Seino, Y., Inomata, S., Sasabe, H., Pu, Y. J. & Kido, J. High-performance green OLEDs using thermally activated delayed fluorescence with a power efficiency of over 100 lm W(-1). *Adv. Mater*. **28**, 2638–2643 (2016).
11. Di, D. et al. High-performance light-emitting diodes based on carbene-metal-amides. *Science* **356**, 159–163 (2017).
12. Aizawa, N. et al. Solution-processed multilayer small-molecule light-emitting devices with high-efficiency white-light emission. *Nat. Commun*. **5**, 5756 (2014).
13. Dai, X. et al. Solution-processed, high-performance light-emitting diodes based on quantum dots. *Nature* **515**, 96–99 (2014).
14. Tan, Z. K. et al. Bright light-emitting diodes based on organometal halide perovskite. *Nat. Nanotechnol*. **9**, 687–692 (2014).
15. Wei, Z. et al. Solution-processed highly bright and durable cesium lead halide perovskite light-emitting diodes. *Nanoscale* **8**, 18021–18026 (2016).
16. Li, X. et al. $CsPbX_3$ quantum dots for lighting and displays: room-temperature synthesis, photoluminescence superiorities, underlying origins and white light-emitting diodes. *Adv. Funct. Mater*. **26**, 2435–2445 (2016).
17. Kim, Y. H. et al. Highly efficient light-emitting diodes of colloidal metal-halide perovskite nanocrystals beyond quantum size. *ACS Nano* **11**, 6586–6593 (2017).
18. Protesescu, L. et al. Nanocrystals of cesium lead halide perovskites (CsPbX(3), X = Cl, Br, and I): novel optoelectronic materials showing bright emission with wide color gamut. *Nano Lett*. **15**, 3692–3696 (2015).
19. Xing, J. et al. High-efficiency light-emitting diodes of organometal halide perovskite amorphous nanoparticles. *ACS Nano* **10**, 6623–6630 (2016).
20. de Mello, J. C., Wittmann, H. F. & Friend, R. H. An improved experimental determination of external photoluminescence quantum efficiency. *Adv. Mater*. **9**, 230–232 (1997).
21. Yang, Y. et al. High-efficiency light-emitting devices based on quantum dots with tailored nanostructures. *Nat. Photon*. **9**, 259 (2015).
22. Lee, S. et al. Growth of nanosized single crystals for efficient perovskite light-emitting diodes. *ACS Nano* **12**, 3417–3423 (2018).
23. Yan, F. et al. Highly efficient visible colloidal lead-halide perovskite nanocrystal light-emitting diodes. *Nano Lett*. **18**, 3157–3164 (2018).
24. Chin, X. Y. et al. Self-assembled hierarchical nanostructured perovskites enable highly efficient LEDs via an energy cascade. *Energy Environ. Sci*. http://dx.doi.org/10.1039/c8ee00293b (2018).
25. Song, J. et al. Room-temperature triple-ligand surface engineering synergistically boosts ink stability, recombination dynamics, and charge injection toward EQE-11.6% perovskite QLEDs. *Adv. Mater*. **30**, e1800764 (2018).
26. Saliba, M. et al. Incorporation of rubidium cations into perovskite solar cells improves photovoltaic performance. *Science* **354**, 206–209 (2016).
27. Chen, W. et al. Efficient and stable large-area perovskite solar cells with inorganic charge extraction layers. *Science* **350**, 944–948 (2015).
28. Zhao, L. et al. Electrical stress influences the efficiency of CH3 NH3 PbI3 perovskite light emitting devices. *Adv. Mater*. **29**, 1605317 (2017).

Acknowledgements This work was supported by the Scientific Research Funds of Huaqiao University (600005-Z16J0038) and the National Natural Science Foundation of China (U1705256). Z.W. thanks Y. Wang (Technical Institute of Physics and Chemistry, Chinese Academy of Sciences) and J. Wang (Institute of Advanced Materials, Nanjing Tech University) for helpful discussions on how to accurately measure the performance of perovskite LED devices. Z.W. also thanks S. Yang (Hong Kong University of Science and Technology) for carrying out time-of-flight/secondary ion mass spectrometry (TOF-SIMS) analysis. Q.X. acknowledges support from the Singapore National Research Foundation through an NRF Investigatorship award (NRF-NRFI2015-03); and from the Ministry of Education via an AcRF Tier2 grant (MOE-2015-T2-1-047) and Tier1 grants (RG 113/16 and RG 194/17). This publication is based in part on work supported by the Canada Research Chairs program, the Natural Sciences and Engineering Research Council of Canada, and the US Department of the Navy, Office of Naval Research (grant N00014-17-1-2524).

Author contributions Z.W., Q.X. and E.H.S. conceived the idea. K.L., J.X. and W.Z. prepared the samples and carried out optical spectroscopy characterizations. L.X. and Y.L. carried out XPS, AFM and some other materials characterizations. K.L., J.L., D.Z., C.Y., W.L. and X.L. fabricated and characterized perovskite LED devices. Z.W. supervised the fabrication and characterization of perovskite LEDs. L.N.Q., X.G., J.K. and F.P.G.A. assisted with device fabrication, measurements and manuscript writing. K.L., Z.W., Q.X. and E.H.S. carried out data analysis and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0575-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0575-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Z.W., E.H.S. or Q.X.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Unless otherwise stated, all chemicals were purchased from Sigma-Aldrich and used as received.

Preparation of perovskite precursor. We first synthesized $CsPbBr_3$ powder and used it as a starting material for precursor preparation. Specifically, we dissolved $PbBr_2$ (10 mmol, 3.67 g) in hydrobromic acid (8 ml), then added CsBr (10 mmol, 2.12 g, dissolved in 3 ml of water) drop by drop, producing an orange precipitate. The precipitate was filtered, washed twice using ethanol, and dried at 60 °C in a vacuum oven for 12 h before use.

CH_3NH_3Br (abbreviated to MABr) was prepared by reacting 12 ml of methylamine (33 wt.% in absolute ethanol) and 11 ml of hydrobromic acid (48 wt.% in H_2O) in an ice bath for 2 h with continuous stirring. The solvent was removed using rotary evaporation at 50 °C to obtain a white MABr powder. For purification, the as-prepared MABr powder was re-dissolved in ethanol and precipitated with diethyl ether. Finally, the white powder was collected by filtration and dried at 60 °C in a vacuum oven for at least 12 h before use.

The $MAPbBr_3$ precursor was prepared by dissolving $PbBr_2$ and MABr (1:1 molar ratio) in DMSO solvent to make a 0.5 M solution. $CsPbBr_3$ powder can be fully dissolved in DMSO solvent to make a 0.5 M starting solution. Then, different amounts of MABr were added to the as-prepared $CsPbBr_3$ solution to make a mixture perovskite precursor, with the mixture being named through the molar ratio of MABr to $CsPbBr_3$. For example, to prepare a precursor of mixture 1.0, 55.98 mg of MABr (0.5 mmol) was added to 1 ml of the as-prepared $CsPbBr_3$ (0.5 mmol) solution.

Fabrication of perovskite LEDs. Prepatterned ITO glasses (20 mm × 20 mm) were ultrasonically washed in, sequentially, detergent solution, deionized water, acetone and ethanol, and then dried with compressed N_2. The substrates were further cleaned with UV-Ozone cleaner (Novascan, PSD) for 30 min before spin-coating. A 40-nm-thick HTL was prepared by spin-coating using PEDOT:PSS (Clevios PV P AI4083) at 4,000 r.p.m. for 60 s and baking at 150 °C for another 15 min. After cooling to room temperature, the substrates were transferred into a nitrogen-filled glove box (H_2O less than 1 part per million (p.p.m.); O_2 less than 1 p.p.m) for deposition of the perovskite layer. The perovskite layer was prepared through the same spin-coating procedures but used a different precursor solution. Specifically, 30 μl of perovskite precursor was dropped onto the substrate and spun at 2,000 r.p.m. for 60 s, during which time (at 30 s) 500 μl of toluene was dropped quickly onto the surface. Another thin layer of PMMA blocking layer was prepared if needed: 50 μl of PMMA solution (0.5 mg ml^{-1} in acetone) was spin-coated onto the as-prepared perovskite layer at 4,000 r.p.m. for 60 s. There was no annealing process and the as-prepared substrates were transferred into a thermal evaporator. The chamber was vacuum-pumped down to 5.0×10^{-4} Pa, and a 40-nm-thick layer of B3PYMPM (Lumtec, Taiwan), 2-nm-thick layer of LiF and 100-nm-thick layer of Al were sequentially evaporated. We defined the area of the perovskite LED device as the area of overlap between the ITO and the Al electrode; it is 3 mm^2 (2 mm × 1.5 mm). We also used a charge-coupled-device (CCD) camera to measure the actual device area.

Material characterization. We characterized surface morphologies by field-emission SEM (using a Hitachi S-8000 scanning electron microscope). To determine the distribution of ions in the perovskite film, we used TOF-SIMS (ION-TOF GmbH, ToF SIMS V). We characterized bright-field, high-angle annular dark field, and element-mapping images with a FEI Talos F200S transmission electron microscope. We prepared the perovskite sample for TEM observations with a focused ion beam (FEI Scios); protective layers of C and Pt were deposited before ion-beam cutting and etching. We recorded XRD patterns using a D8 Advance diffractometer (Bruker AXS). Ultraviolet/visible and steady-state photoluminescence spectra were acquired using a Flame spectrometer (Ocean Optics) in a glovebox. We measured PLQYs using a blue excitation laser (405 nm), an integrating sphere and a Flame spectrometer. Photoluminescence decay curves were measured using a fluorescence lifetime imaging microscope (FLIM, Leca TCS SP8) with a pulsed excitation laser of 405 nm.

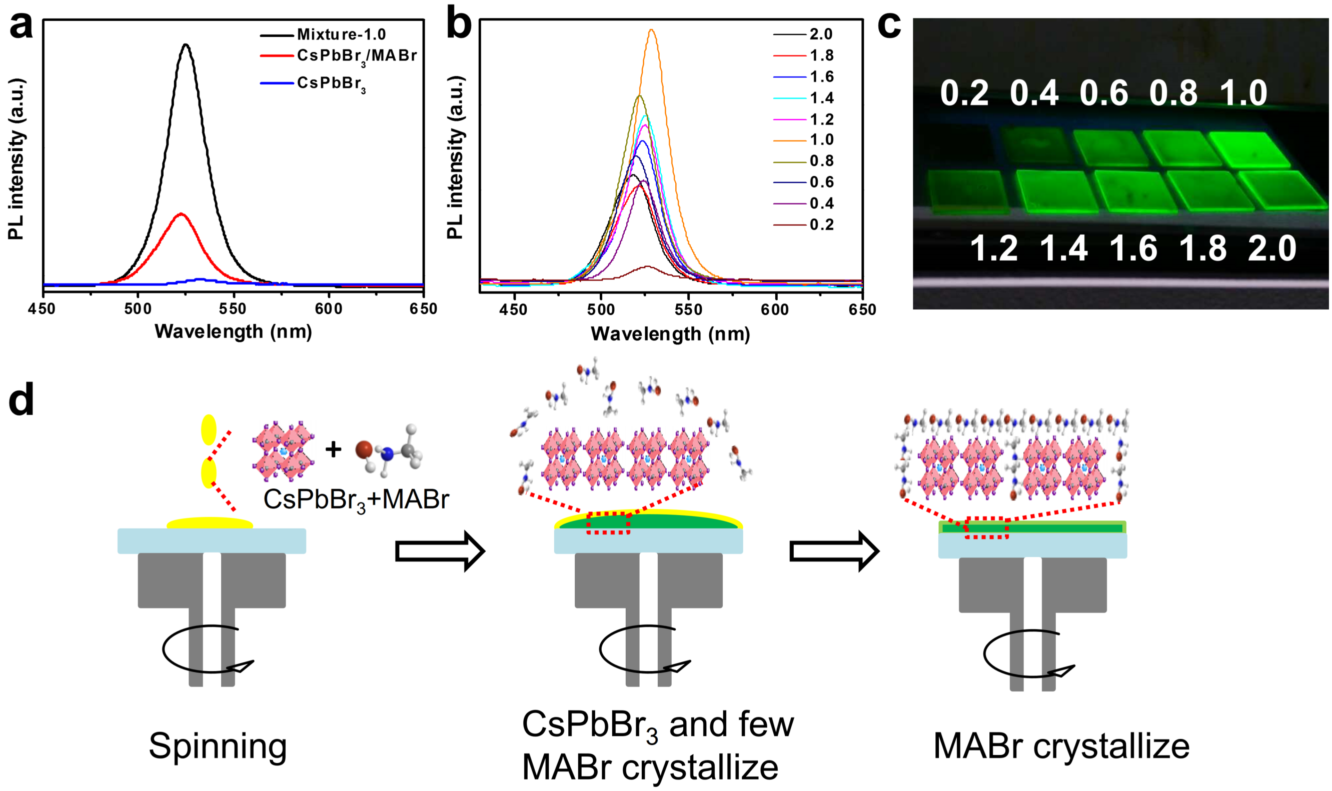
Performance evaluation of perovskite LEDs. We inserted the as-fabricated perovskite LED devices into a home-made test socket and took measurements in a glovebox (Supplementary Information, videos S2 and S3, from which we can see that the luminance loss caused by light absorbance and reflection of the glove box glass is 16%). Using a Keithley 2400 instrument, we measured J–V data from 0 V to 5 V with a step voltage of 0.2 V and delay time of 3 s; simultaneously, we measured the luminance using a luminance meter (Konica Minolta, LS-160 or CS-200). Electroluminescence characteristics were recorded with a Flame spectrometer (Ocean Optic). The current efficiency was calculated by dividing the luminance by the current density. The EQE was calculated using Lambertian emission profiles and the obtained electroluminescence spectra. The initial high-performance device with an EQE of 16% was first investigated by the Nanyang Technological University group; the electroluminscence characteristics were further studied there as well.

Measurement of device lifetime by accelerated ageing. Prior studies have shown that the product of the initial luminance (L_0) of the lifetime measurement and the T_{50} lifetime (defined as the time when the luminance drops to 50% of L_0) is a constant: $L_0^n \times T_{50} =$ constant, where the n is the acceleration factor. The n factor can be determined experimentally by running lifetime tests for different L_0 values. The equation can be rewritten in the form: $\log T_{50} = K - n\log L_0$. In this way, n is obtained as the slope of the linear fitting curve of the various measured T_{50} and L_0 values.

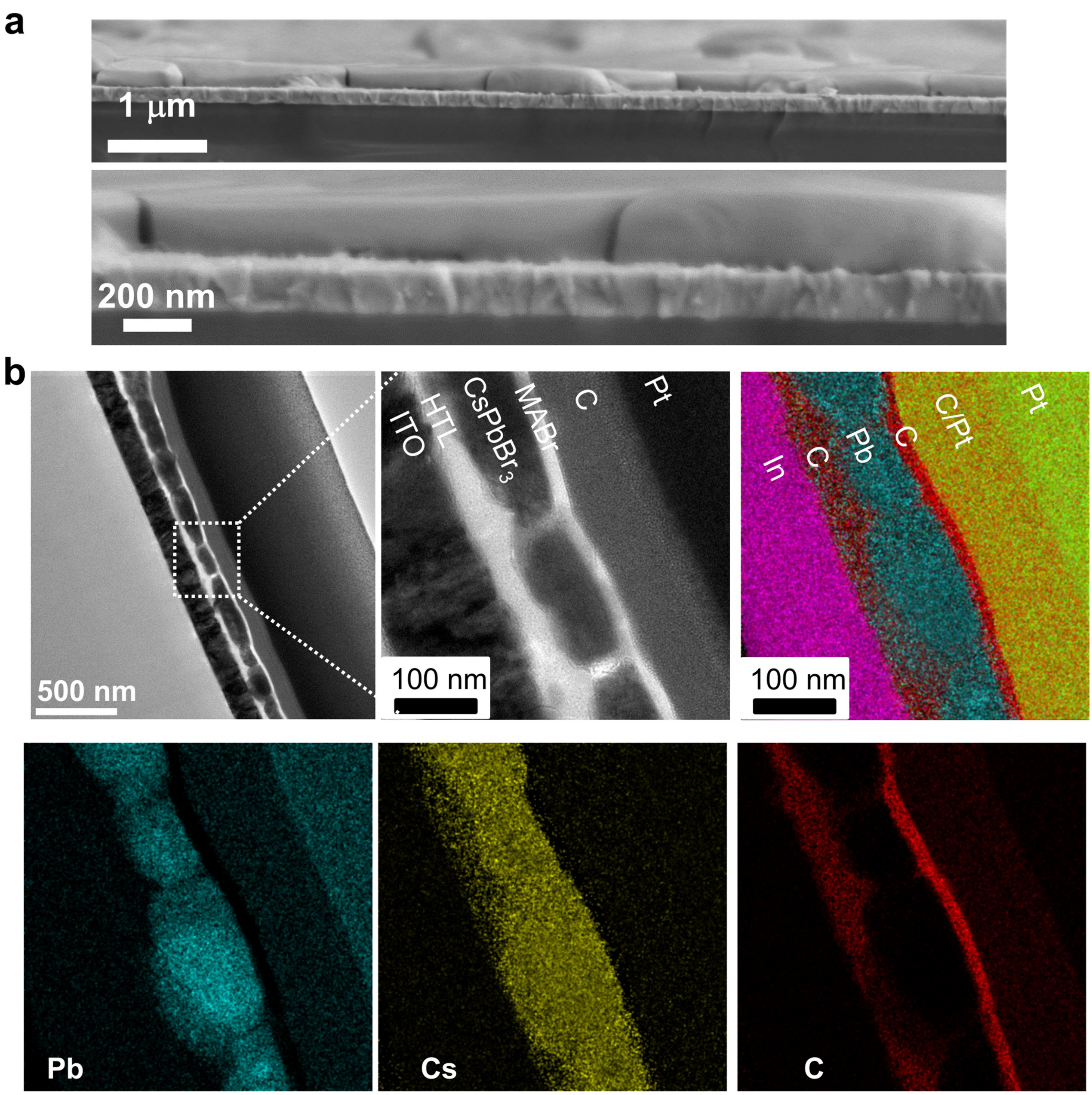
Measurement of operational lifetime (constant luminance). As shown in Supplementary Information, video S4, we could maintain the luminance of our perovskite LED devices at around 100 cd cm^{-2} by carefully tuning the applied current. In other words, we had to increase the applied current a little bit once an obvious decrease in luminance was observed. The device worked steadily for around 46 h, and then began to degrade rapidly.

Data availability

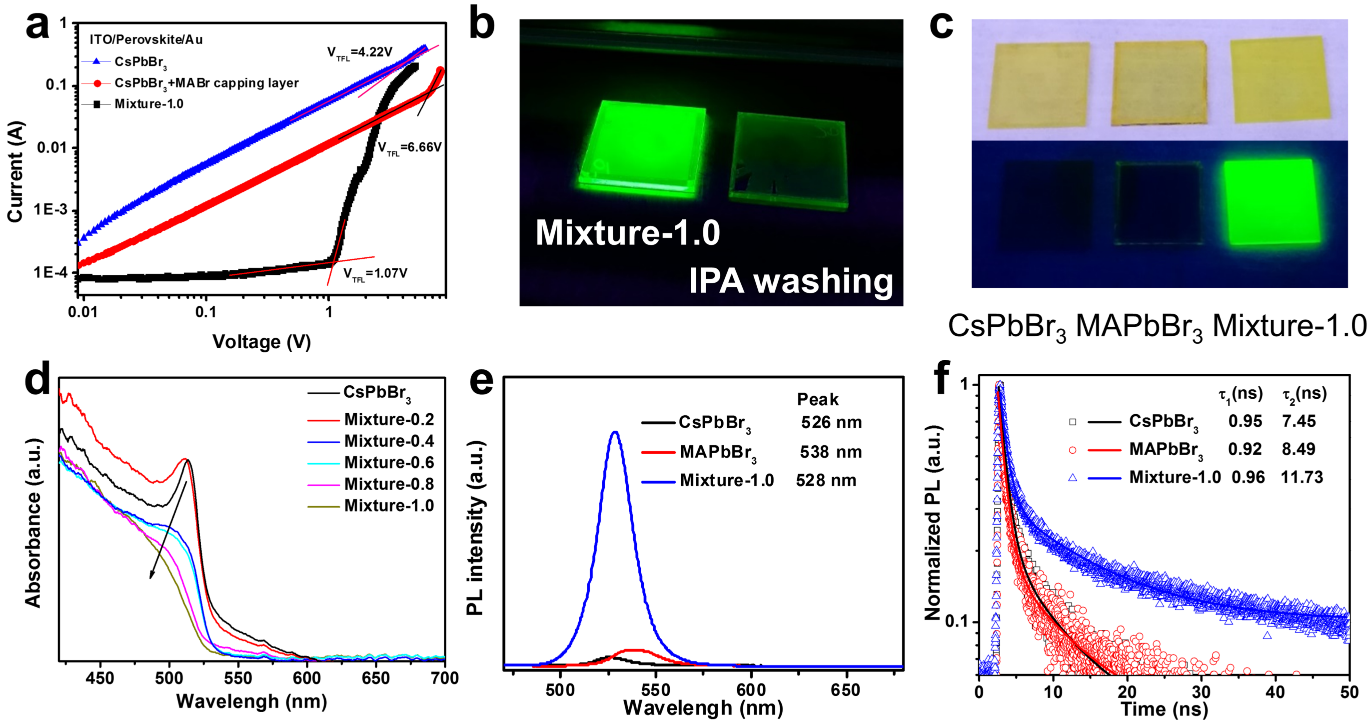
The data that support the findings of this study are available from the corresponding author upon reasonable request.



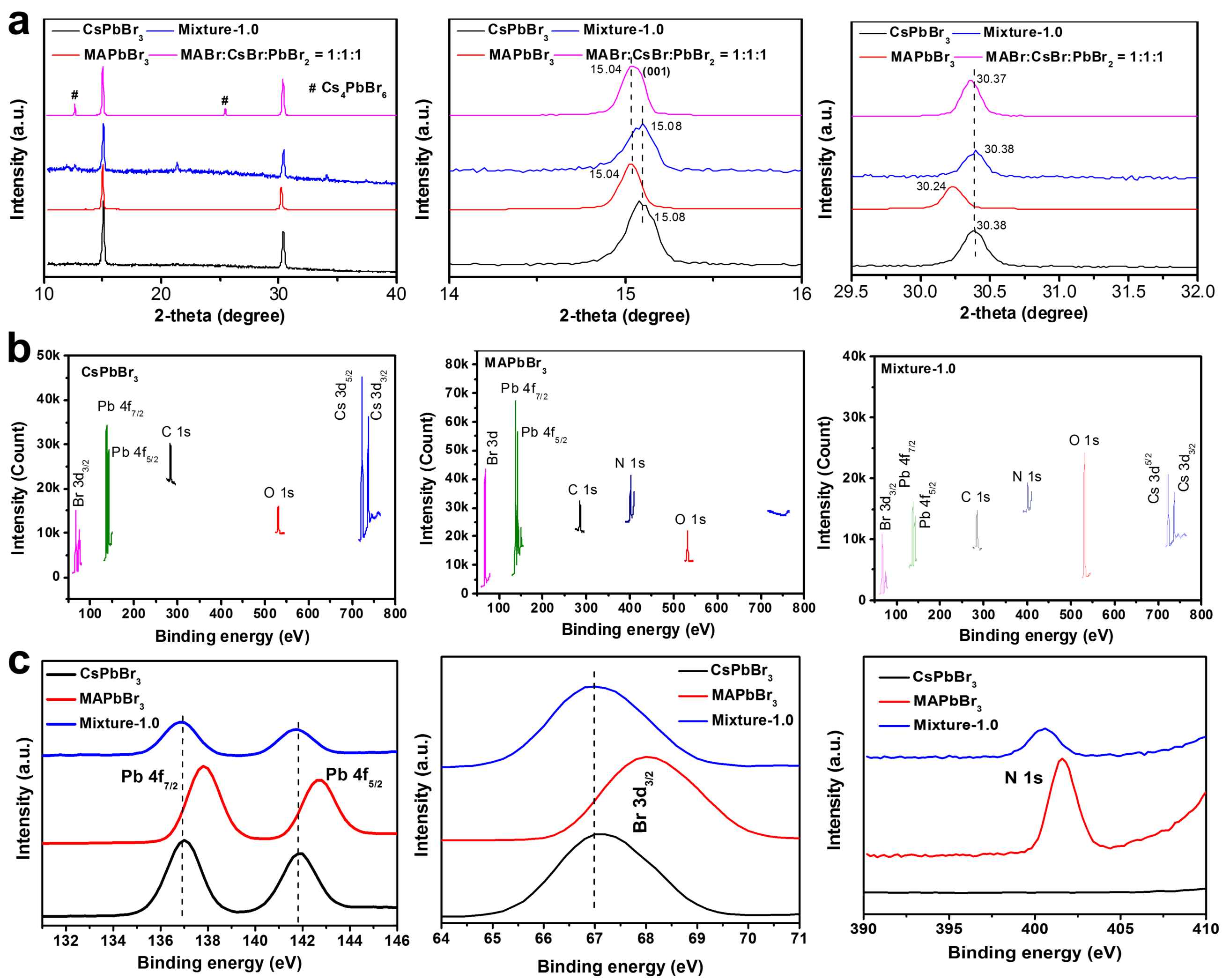
Extended Data Fig. 1 | Emission properties and formation mechanism of mixture perovskites. a, Photoluminescence spectra of single-layered $CsPbBr_3$, bilayered $CsPbBr_3$/MABr, and quasi-core/shell-structured mixture-1.0 films. **b**, Photoluminescence spectra of various mixture perovskite films with different amounts of MABr. The numbers in the key indicate the molar ratio of MABr to $CsPbBr_3$. **c**, Photograph of mixture perovskite films under ultraviolet light. **d**, The formation process of the $CsPbBr_3$/MABr quasi-core/shell structure. The yellow areas in the left panel denote the precursor solution.



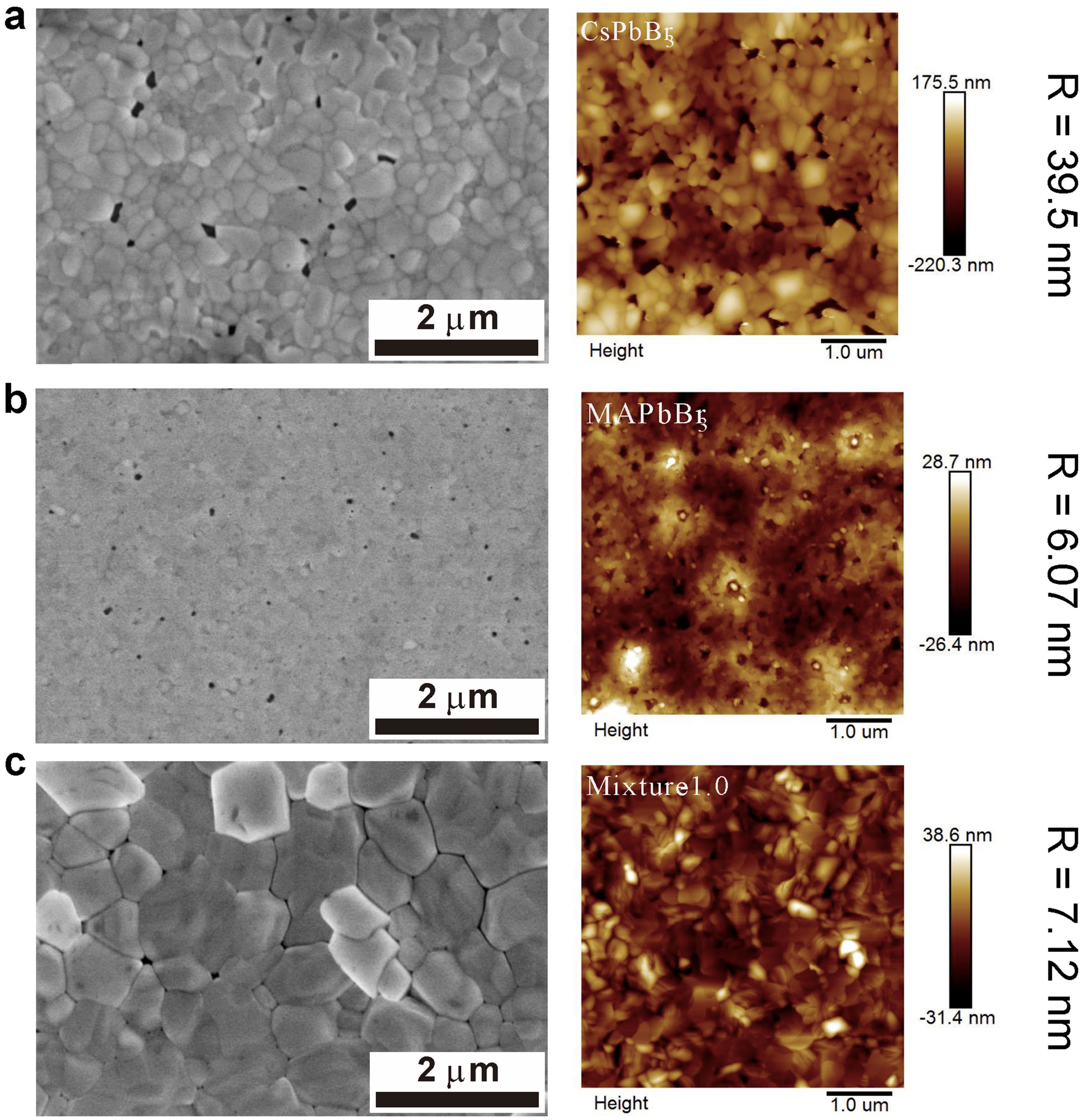
Extended Data Fig. 2 | SEM and TEM images of the mixture-1.0 films. **a**, Cross-sectional SEM images of the as-prepared mixture-1.0 film at different magnifications. **b**, Cross-sectional TEM images (greyscale) and elemental mapping images (in colour) of the mixture-1.0 film. The sample was prepared using a focused ion beam, and the top C and Pt layers were predeposited to protect the perovskite film.



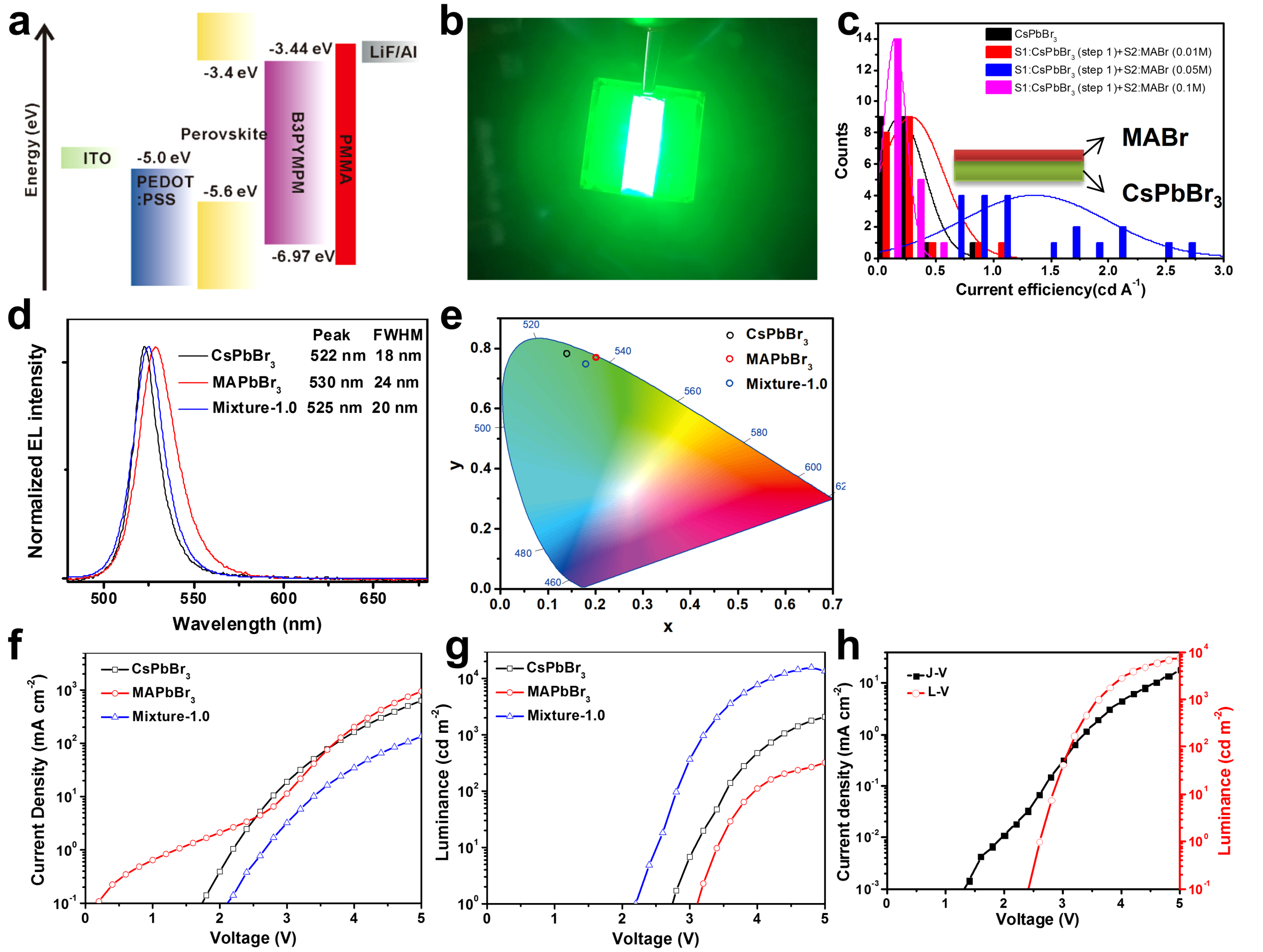
Extended Data Fig. 3 | Comparison of $CsPbBr_3$, $MAPbBr_3$ and mixture-1.0 perovskite films. **a**, Estimation of trap-state density using the space-charge-limited current method, obtained using dark *I*–*V* curves of perovskite devices with the structure ITO/perovskite/Au. V_{TFL}, trap-filled limiting voltage. **b**, Photographs of a pristine mixture-1.0 perovskite film (left) and a film treated by washing with IPA (right; to remove the MABr capping layer) under ultraviolet light. **c**, Photograph of the three perovskite films under ultraviolet light. **d**, Ultraviolet–visible absorbance spectra of the $CsPbBr_3$ film and different mixture films. The black arrows indicate the disappearance of the exciton peak from the $CsPbBr_3$ sample to the mixture-1.0 sample. **e**, **f**, Photoluminescence spectra (**e**) and time-resolved photoluminescence decay curves (**f**; excitation source: 400 nm, 4 μW) of the $CsPbBr_3$, $MAPbBr_3$ and mixture-1.0 films. τ is the lifetime.



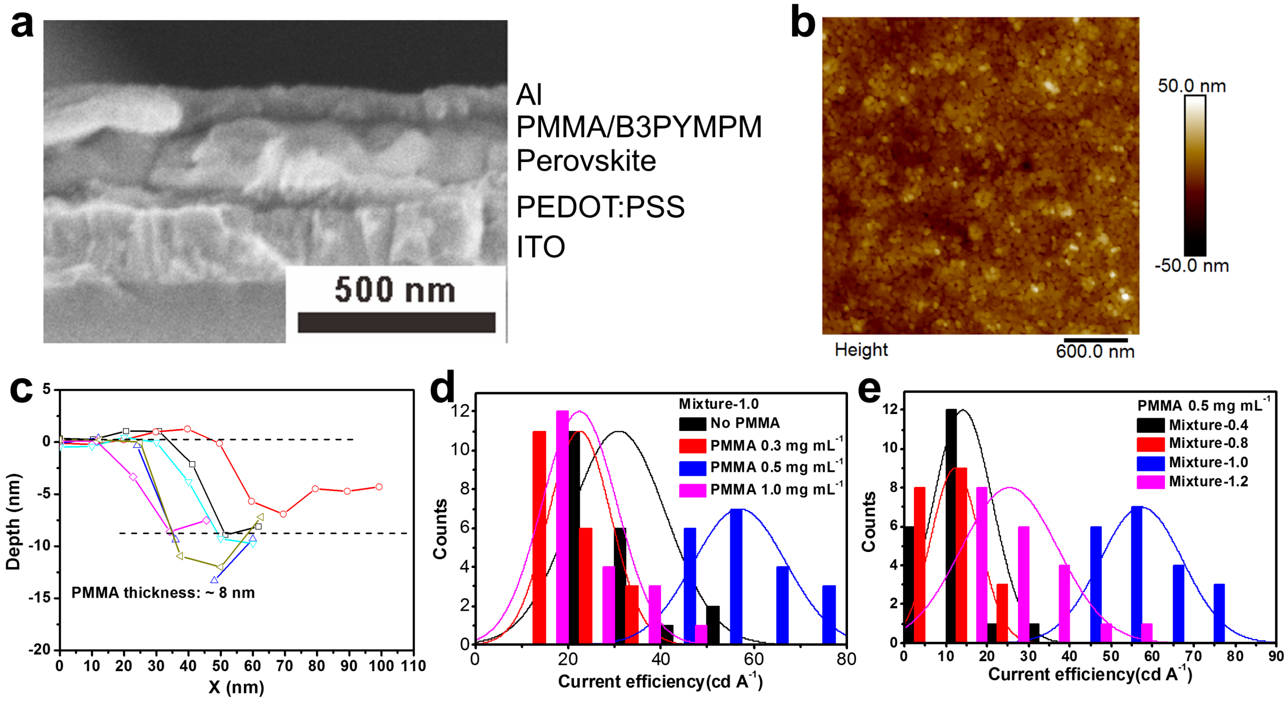
Extended Data Fig. 4 | XRD and XPS characteristics of the perovskite films. a, Original and magnified XRD patterns from $CsPbBr_3$, $MAPbBr_3$, mixture-1.0, and MABr + $PbBr_2$ + CsBr mixture perovskite films. 2-theta is the diffraction angle multiplied by 2. **b**, **c**, XPS results from $CsPbBr_3$, $MAPbBr_3$ and mixture-1.0 perovskite films, indicating that there are $CsPbBr_3$ and MABr phases in the mixture-1.0 film, but not in $MAPbBr_3$.



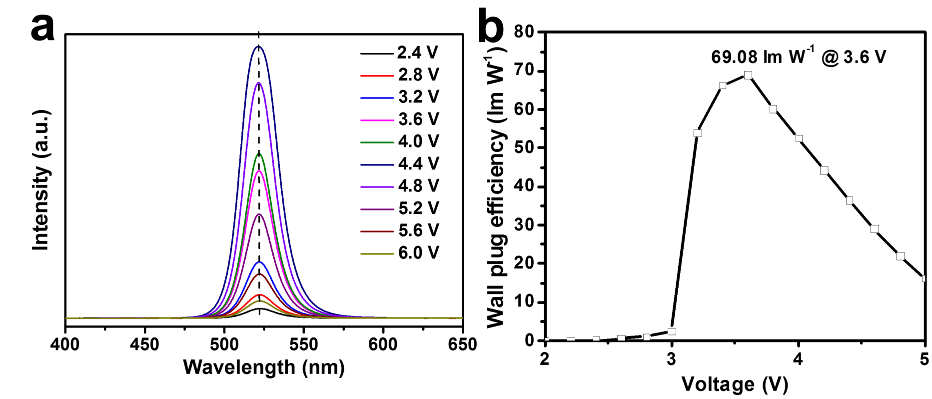
Extended Data Fig. 5 | Comparison of the morphology of the three perovskite films. Left, top-view SEM image; centre, AFM topography; and right, root-mean-square roughness (*R*) of the **a**, $CsPbBr_3$, **b**, $MAPbBr_3$ and **c**, mixture-1.0 perovskite films.



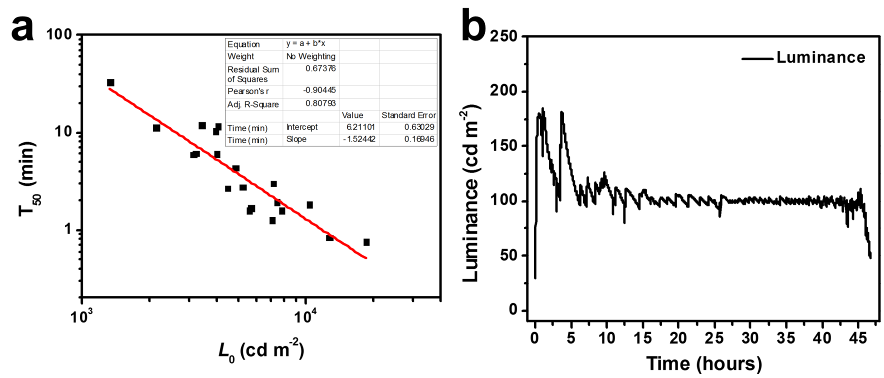
Extended Data Fig. 6 | Fabrication and evaluation of a perovskite LED device. a, Energy-level diagram of the as-fabricated perovskite LED. **b**, Photograph of a large-area perovskite LED device (6 mm × 20 mm). **c**, Device performance of the bilayered $CsPbBr_3$/MABr perovskite LEDs. S1 and S2 are step 1 and step 2. **d**, **e**, Electroluminescence spectra (**d**) and CIE chromatic diagram (**e**) of the three as-fabricated perovskite LEDs. **f**, **g**, *J*–*V* (**f**) and *L*–*V* (**g**) curves of the three as-fabricated three perovskite LEDs. **h**, *L*–*V* and *J*–*V* curves of the best-performing mixture-1.0 perovskite LED without a PMMA layer.



Extended Data Fig. 7 | Performance optimization of the mixture-1.0 perovskite LED. a, Cross-sectional SEM image of a perovskite LED device with a PMMA blocking layer. **b**, **c**, AFM topography image (**b**) and thickness measurement (**c**) of the optimized PMMA layer. The differently coloured lines show multiple test results from the same sample. **d**, **e**, Optimizing device performance by tuning the thickness of the PMMA layer (**d**) and the amount of MABr additives in the perovskite precursor (**e**).



Extended Data Fig. 8 | Electroluminescence spectra and power efficiency of the best-performing mixture-1.0 perovskite LED device. **a**, Electroluminescence spectra at various applied voltages. **b**, Power efficiency curve.



Extended Data Fig. 9 | Measurement of operational lifetime of typical mixture-1.0 devices. **a**, Extrapolation of lifetime from accelerated ageing tests, from which one can see that the acceleration factor, *n*, is about 1.5 in the devices investigated herein. **b**, Measurement of the operational lifetime of a perovskite LED working in continuous luminance mode; by carefully tuning the applied current, we could maintain a luminance output of around 100 cd m^{-2}.

Extended Data Table 1 | Stability performance of other reported highly efficient green perovskite LEDs (with EQEs of more than 10%)

Articles	Max. EQE	Emitting materials	Stability performance
Ref. 7	10.4%	$Cs_{0.87}MA_{0.13}PbBr_3$	V = 3.7 V, L_0 = ~610 cd m^{-2}, T_{50} = 40 s; T_{50} at 100 cd m^{-2} is determined to be 10 min.
Ref. 22	12.1%	$MAPbBr_3$	J = 5 mA cm^{-2}, L_0: not indicated, T_{50} = 135 min.
Ref. 9	14.36%	$PEA_2(FAPbBr_3)_{n-1}PbBr_4$	J = 0.5 mA cm^{-2}, L_0 = 270 cd m^{-2}, T_{50} = 65 min; T_{50} at 100 cd m^{-2} is determined to be 4.8 h.
Ref. 23	12.9%	$MAPbBr_3$	J = 0.3 mA cm^{-2}, L_0 = 100 cd m^{-2}, T_{50} = 6 min
Ref. 24	13.4%	$(OA)_2(FA)_{n-1}Pb_nBr_{3n+1}$ and $FAPbBr_3$	J = 0.36 mA cm^{-2}, L_0 = 105 cd m^{-2}, T_{50} = 800s
Ref. 25	11.6%	FA-doped $CsPbBr_3$	Not mentioned
This work	20.3%	$CsPbBr_3$@MABr	J = 166.67 mA cm^{-2}, L_0 = 7130 cd m^{-2}, T_{50} = 10.42 min; T_{50} at 100 cd m^{-2} is determined to be 104.56 h; Lifetime measured in continual mode with L of 100 cd m^{-2} is ~46 h;

V is the driving voltage; J is the applied current density; L_0 is the initial luminance; T_{50} is the time over which the luminance decreases to 50% of L_0. T_{50} at 100 cd m^{-2} is calculated using: $L_0^n T_{50}$= constant. We assume that the acceleration factor, n, is 1.5.

https://doi.org/10.1038/s41586-018-0576-2

Perovskite light-emitting diodes based on spontaneously formed submicrometre-scale structures

Yu Cao[1,7], Nana Wang[1,7], He Tian[2,7], Jingshu Guo[3,7], Yingqiang Wei[1], Hong Chen[1], Yanfeng Miao[1], Wei Zou[1], Kang Pan[1], Yarong He[1], Hui Cao[1], You Ke[1], Mengmeng Xu[1], Ying Wang[1], Ming Yang[1], Kai Du[2], Zewu Fu[1], Decheng Kong[1], Daoxin Dai[3], Yizheng Jin[4], Gongqiang Li[1], Hai Li[1], Qiming Peng[1], Jianpu Wang[1]* & Wei Huang[1,5,6]*

Light-emitting diodes (LEDs), which convert electricity to light, are widely used in modern society—for example, in lighting, flat-panel displays, medical devices and many other situations. Generally, the efficiency of LEDs is limited by nonradiative recombination (whereby charge carriers recombine without releasing photons) and light trapping[1–3]. In planar LEDs, such as organic LEDs, around 70 to 80 per cent of the light generated from the emitters is trapped in the device[4,5], leaving considerable opportunity for improvements in efficiency. Many methods, including the use of diffraction gratings, low-index grids and buckling patterns, have been used to extract the light trapped in LEDs[6–9]. However, these methods usually involve complicated fabrication processes and can distort the light-output spectrum and directionality[6,7]. Here we demonstrate efficient and high-brightness electroluminescence from solution-processed perovskites that spontaneously form submicrometre-scale structures, which can efficiently extract light from the device and retain wavelength- and viewing-angle-independent electroluminescence. These perovskites are formed simply by introducing amino-acid additives into the perovskite precursor solutions. Moreover, the additives can effectively passivate perovskite surface defects and reduce nonradiative recombination. Perovskite LEDs with a peak external quantum efficiency of 20.7 per cent (at a current density of 18 milliamperes per square centimetre) and an energy-conversion efficiency of 12 per cent (at a high current density of 100 milliamperes per square centimetre) can be achieved—values that approach those of the best-performing organic LEDs.

Organometal halide perovskites are promising light-emitting materials for solution-processed LED applications because of their high photoluminescence quantum efficiency (PLQE), good charge mobility and excellent colour purity[10–12]. In order to achieve high-efficiency perovskite LEDs, extensive efforts have been made to reduce the nonradiative recombination and improve the PLQE[13–15]. So far, the PLQE of perovskite film has reached as much as 70%, but the peak external quantum efficiency (EQE) of the device electroluminescence is still less than 15%[15–17]. Studies of device physics have shown that, in principle, charge balance is not a limiting factor for the perovskite LED[17]. Therefore, the main loss of efficiency must be due to light trapping—a general efficiency-limiting factor in most types of LEDs. In perovskite LEDs, light trapping could be more serious than in organic LEDs, because the refractive index of perovskites is much higher than those of organic materials[18].

Here we demonstrate effective extraction of trapped light from perovskite LEDs by a spontaneously formed perovskite submicrometre-scale structure. The fabrication process is shown schematically in Fig. 1a. A precursor solution of 5-aminovaleric acid (5AVA), formamidinium iodide (FAI) and PbI_2 with a molar ratio of 0.7/2.4/1 dissolved in *N*,*N*-dimethylformamide (DMF; 7 wt.%) was used to deposit perovskite films, which were annealed at 100 °C for 16 min (see Methods for details) before depositing the top charge-transport layer. The device structure is indium tin oxide (ITO)/polyethylenimine ethoxylated (PEIE)-modified zinc oxide (ZnO; thickness 30 nm)/perovskite (around 50 nm)/poly(9,9-dioctyl-fluorene-co-*N*-(4-butylphenyl)diphenylamine) (TFB; 40 nm)/molybdenum oxide (MoO_x; 7 nm)/gold (Au; 60 nm). A cross-sectional scanning transmission electron microscope (STEM) image (Fig. 1b) shows the formation of discrete submicrometre-structured perovskites in the emitting layer. Scanning electron microscope (SEM) observations (Fig. 1c) further show that the perovskites are faceted platelets with roughly rectangular shapes. The platelets are randomly tiled on the substrate, and the size of the platelets is between 100 nm and 500 nm. Optical microscope, SEM and atomic force microscopy (AFM) images of different magnifications show that the submicrometre structure is homogeneously distributed on the whole substrate (Extended Data Fig. 1). We note that the perovskite submicron platelets are formed directly through spin-coating the precursor solution, unlike in widely reported methods where organic ligands are used to synthesize perovskite nanocrystals[19,20].

High-angle annular dark-field (HAADF)-STEM tomography observations show that the perovskite submicrometre platelets are embedded in a roughly 8-nm-thick organic layer. In order to avoid possible interference of the TFB layer during the STEM measurement, we prepared a sample by depositing a gold layer directly on top of the ZnO-PEIE/perovskite. A cross-section STEM tomography reconstruction from a series of images of a tilted sample (Fig. 1d) shows a single layer of perovskite submicrometre platelets distributed on top of the ZnO-PEIE layer, consistent with Fig. 1b. Figure 1e is a STEM tomography image at higher magnification. The contrast seen in this image suggests the existence of a thin layer in which atoms of low atomic number fill in the gaps between the perovskite platelets. Associated energy-dispersive X-ray spectroscopy (EDS) measurements show that carbon, but not lead, has accumulated in the thin layer (which is around 8 nm thick) (Fig. 1f, g). Therefore, we can conclude that there is a thin organic layer filling in the gaps between the perovskite platelets, and that no organic layer can be observed underneath or above the perovskite submicrometre platelets (Fig. 1e–g).

We next investigated how this submicrometre-scale structure is formed. After the 5AVA-perovskite precursor solution has been

[1]Key Laboratory of Flexible Electronics (KLOFE) and Institute of Advanced Materials (IAM), Jiangsu National Synergetic Innovation Center for Advanced Materials (SICAM), Nanjing Tech University (NanjingTech), Nanjing, China. [2]Center of Electron Microscope, State Key Laboratory of Silicon Material, School of Material Science and Engineering, Zhejiang University, Hangzhou, China. [3]Centre for Optical and Electromagnetic Research, State Key Laboratory for Modern Optical Instrumentation, Zhejiang Provincial Key Laboratory for Sensing Technologies, Zhejiang University, Hangzhou, China. [4]Center for Chemistry of High-Performance and Novel Materials, State Key Laboratory of Silicon Materials, and Department of Chemistry, Zhejiang University, Hangzhou, China. [5]Key Laboratory for Organic Electronics and Information Displays, Institute of Advanced Materials, Nanjing University of Posts and Telecommunications, Nanjing, China. [6]Shaanxi Institute of Flexible Electronics (SIFE), Northwestern Polytechnical University (NPU), Xi'an, China. [7]These authors contributed equally: Yu Cao, Nana Wang, He Tian, Jingshu Guo. *e-mail: iamjpwang@njtech.edu.cn; iamwhuang@nwpu.edu.cn

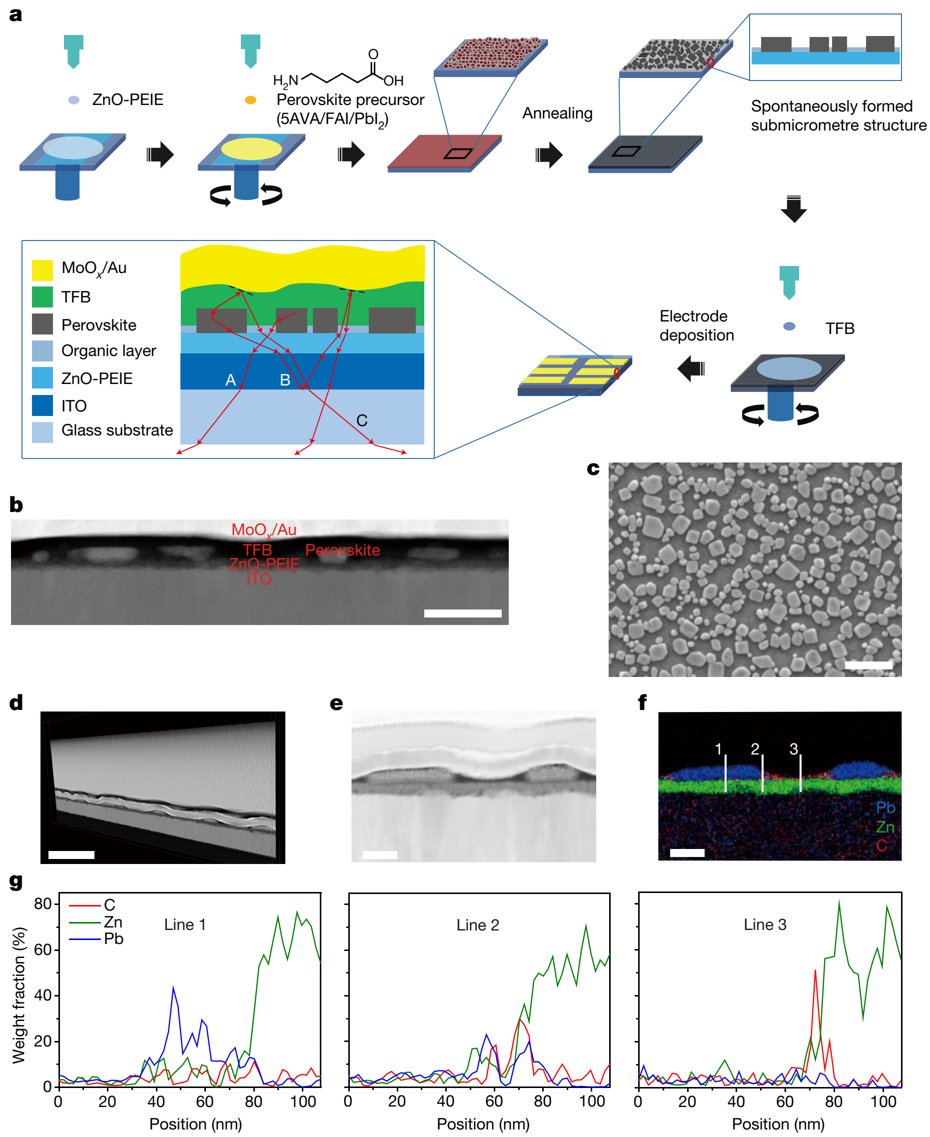


Fig. 1 | Device fabrication and formation of submicrometre structure. **a**, Fabrication of the device and formation of submicrometre structure. Rays A, B and C, which represent light trapped in devices with a continuous emitting layer, can be extracted by the submicrometre structure. **b**, STEM image of the fabricated device. The scale bar represents 200 nm. **c**, SEM image of the perovskite. The scale bar represents1 µm. **d**, Tomographic slice of the HAADF-STEM reconstruction of a sample with the structure ITO/ZnO-PEIE/perovskite/Au. The scale bar represents 500 nm. **e**, Cross-section HAADF-STEM tomography image at high magnification. The scale bar represents 100 nm. **f**, Corresponding EDS composite map. The scale bar represents 100 nm. **g**, EDS profiles of carbon, zinc and lead derived from the lines indicated in **f**.

spin-coated onto the ZnO-PEIE substrate, X-ray diffraction (XRD) measurements show the formation of α-phase crystalline $FAPbI_3$ perovskite (Extended Data Fig. 2). By using Scherrer's equation, we can estimate the size of crystallites to be around 40 nm, which is consistent with the SEM observations (Extended Data Fig. 2). Upon annealing, the perovskite crystallites grow and become submicrometre-sized platelets. Meanwhile, the temporal evolution of grazing-angle reflectance Fourier transform infrared spectroscopy (FTIR) spectra (Extended Data Fig. 3) shows that the O–H stretching vibration of 5AVA decreases while the peaks of amide I and amide II become evident as the annealing time increases. This result suggests that 5AVA undergoes a dehydration reaction on top of the ZnO-PEIE, leading to the formation of an organic layer between the perovskite submicrometre platelets. We believe that the thin organic insulating layer resulting from the dehydration reaction of 5AVA can prevent LED leakage currents caused by low coverage of the perovskite layer[21].

The concave–convex structure formed by high-index perovskite and low-index organics (together with the subsequently deposited TFB layer) can extract wide-angle light trapped in the waveguide modes, which is often achieved by embedding insulating grids within the organic layers in organic LEDs[8]. As shown in the enlarged view in Fig. 1a, the wide-angle light (ray A) can enter the low-index organic layer and propagate into the glass substrate. Moreover, the formed perovskite submicrometre platelets have a flat top surface and a relatively uniform height distribution, as confirmed by STEM tomography (Fig. 1d) and AFM measurements (Fig. 2). We emphasize that the discrete perovskite platelets can greatly affect the morphology of the subsequently deposited films. The TFB layer has similar spontaneously formed submicrometre structures, which further carry over to the MoO_x/Au electrode, resulting in a corrugated metal thin film with a depth of about 30 nm (Fig. 2). These randomly distributed submicrometre structures can extract light from the waveguide mode in all directions (indicated by rays B and C in Fig. 1a), and will not introduce any spectrum shift and angular dependence.

To verify the quality of perovskite films with this unique submicrometre structure, we measured their optical properties. The absorption and photoluminescence spectra (Fig. 3a) show features typical of three-dimensional (3D) $FAPbI_3$, with absorption until 830 nm

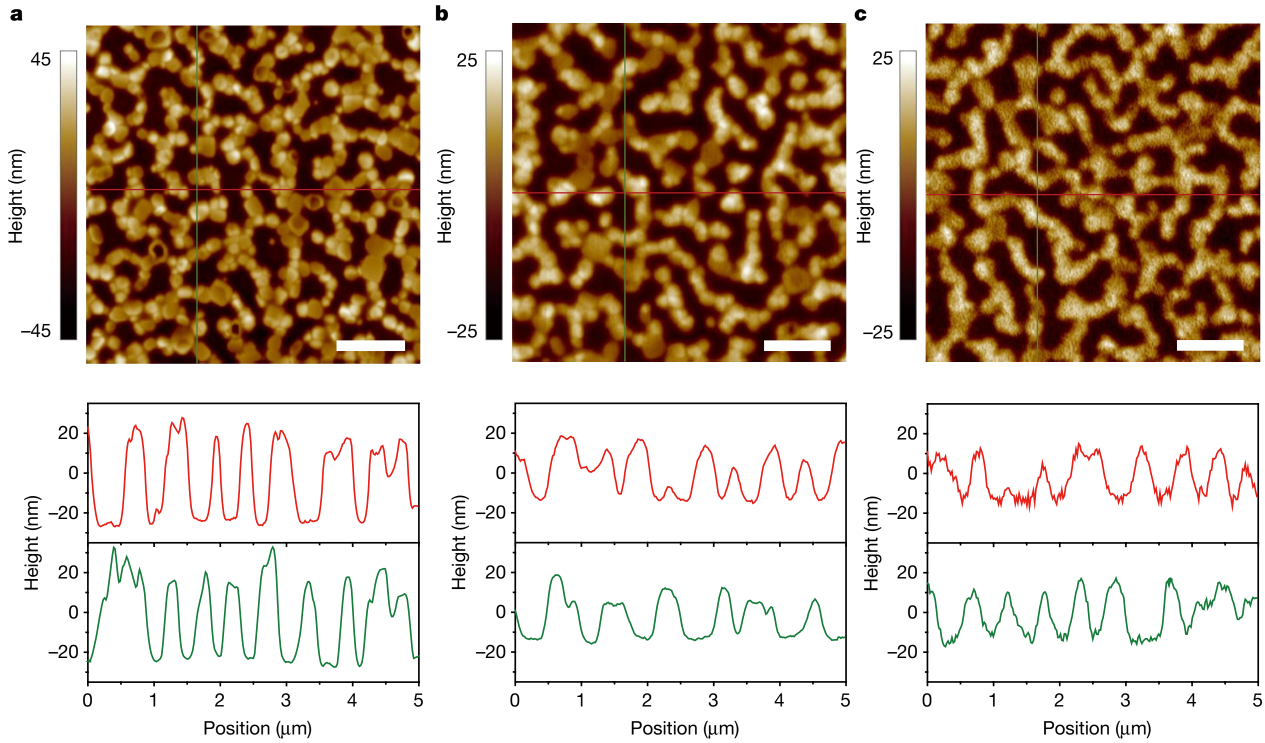


Fig. 2 | AFM height images and multiple line scans. The scale bar represents 1 μm. **a**, ZnO-PEIE/perovskite. The perovskite submicrometre platelets have a height of around 40–50 nm. **b**, ZnO-PEIE/perovskite/TFB. This structure duplicates the morphology of perovskite submicrometre platelets shown in **a**. **c**, ZnO-PEIE/perovskite/TFB/MoO_x/Au. This structure also duplicates the morphology of the perovskite layer.

and a photoluminescence peak at 800 nm[13]. The photoluminescence spectrum has a line width of 75 meV, which is much narrower than those of reported $FAPbI_3$ films[13,22], indicating a highly ordered structure in our perovskites. XRD data (Fig. 3b) show that the film has good crystallinity with peaks at around 14°, 28°, 42° and 56°, corresponding to the (111), (222), (333) and (444) crystal planes, respectively, of α-phase $FAPbI_3$ (ref. [23]). The sharp and strong diffraction peaks suggest that the perovskite submicrometre platelets are highly oriented in the direction perpendicular to the substrate, which is consistent with the SEM measurements (Fig. 1c). The perovskite film has a high PLQE of up to about 70% (Fig. 3c), and the PLQE is maintained at a high level of more than 50% at an excitation intensity as low as 0.1 mW cm^{-2}. The results of PLQE measurements suggest that trap-assisted nonradiative recombination—which is a general limiting factor in 3D perovskites[10,12]—is not substantial in our perovskite platelets. Low trap-assisted nonradiative recombination is confirmed by transient photoluminescence decay measurements, which show a very long photoluminescence lifetime (of about 6 μs) at a low carrier density of 3.4×10^{13} cm^{-3} (Fig. 3d). Transient photoluminescence measurements under various excitation intensities show that a transition from trap-assisted recombination to bimolecular recombination occurs at a carrier density of about 10^{13} cm^{-3}. By fitting the data using a generic kinetic model (Fig. 3d)[24], we can obtain a trap density of 1.5×10^{13} cm^{-3}. Notably, this value is more than one order of magnitude lower than that of previously reported $FAPbI_3$ perovskite films (more than 9×10^{14} cm^{-3})[25]. Extended Data Fig. 4 shows that, without 5AVA in the precursor solution, irregular perovskite clusters are formed with poor crystallinity, strong trap-assisted recombination and very low

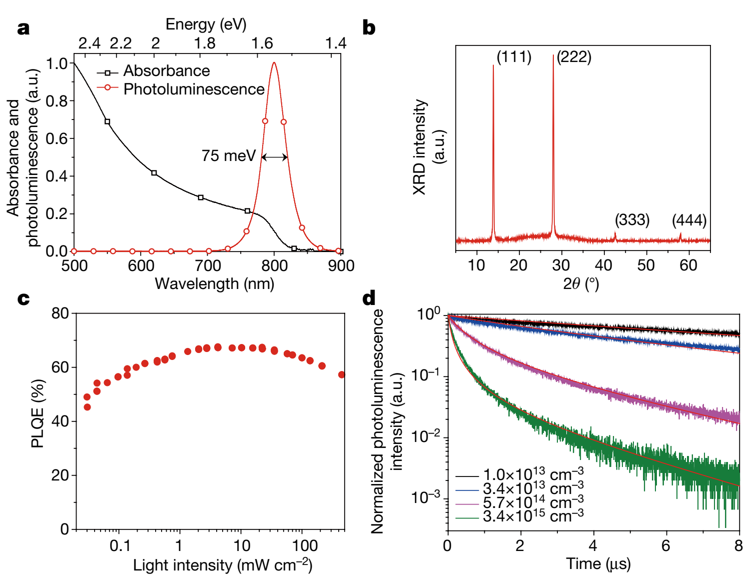


Fig. 3 | Properties of submicrometre-structured perovskite films. a, Absorption and photoluminescence spectra of our perovskite on a ZnO-PEIE substrate. **b**, XRD data from the perovskite films show features of α-phase $FAPbI_3$ with good crystallinity, highly oriented in the perpendicular direction. **c**, Excitation-intensity-dependent PLQE results show a high PLQE of up to 70%; the PLQE is greater than 50% even when the excitation energy is as low as 0.1 mW cm^{-2}, suggesting low trap-assisted nonradiative recombination. **d**, Time-resolved photoluminescence decay transients of submicrometre-structured perovskite under different excitation intensities. Solid lines are fits from the generic kinetic model[24], and a low trap density of 1.5×10^{13} cm^{-3} can be obtained.

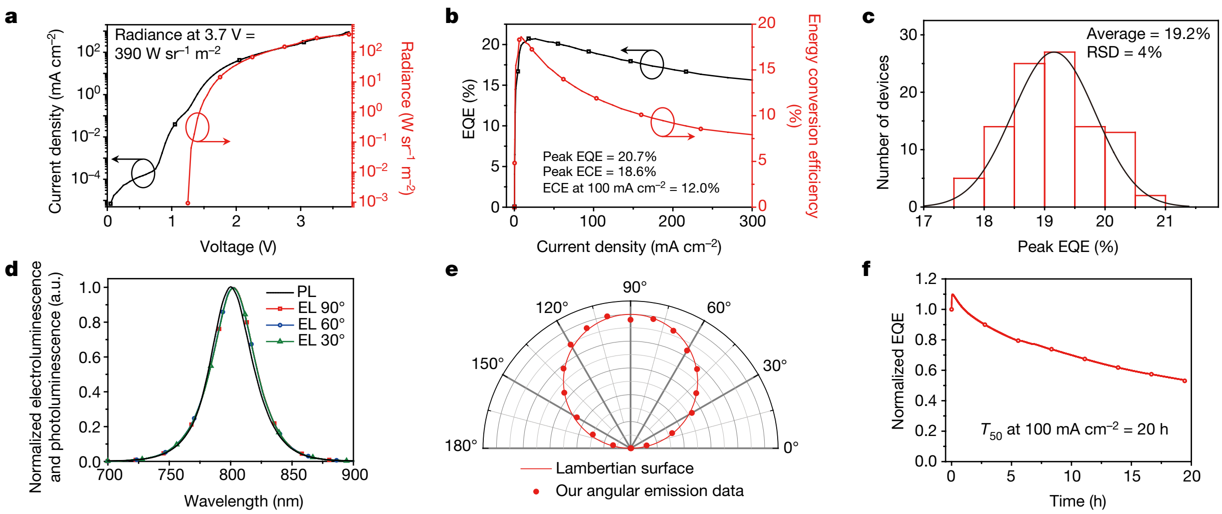


Fig. 4 | Optoelectronic characteristics of our perovskite LEDs. **a**, Dependence of current density and radiance on the voltage. **b**, EQE and ECE plotted against current density. A peak EQE of 20.7% was achieved under a current density of 18 mA cm^{-2}. **c**, Histogram of peak EQEs. Statistics from 100 devices show an average peak EQE of 19.2% with a relative standard deviation (RSD) of 4%. **d**, Photoluminescence (PL) and electroluminescence (EL) spectra of the device at different viewing angles. **e**, The angular distribution of the radiation intensity follows a Lambertian profile. **f**, Stability of the device measured at a constant current density of 100 mA cm^{-2}.

PLQEs at low excitation intensities. These findings indicate that 5AVA has the important role of passivating surface defects and enhancing the emission properties of perovskites. We note that a similar effect has previously been observed[26–29].

These analyses suggest that our submicrometre-scale crystalline perovskites have great potential for realizing high-efficiency LED devices. The current-density/radiance/voltage characteristics of LEDs based on the structured perovskites are shown in Fig. 4a. Owing to the good charge mobility of 3D perovskites (in comparison with organic emitters) and efficient charge injection, the current density and radiance increase quickly once the device turns on at 1.25 V, yielding a high brightness of up to 390 W sr^{-1} m^{-2} at a low voltage of 3.7 V. The device current is low before the electroluminescence turns on, indicating that the leakage current is not important in our LED devices, and further suggesting that the insulator layer between the perovskite submicrometre platelets is pinhole free. The peak EQE reaches 20.7% at a current density of 18 mA cm^{-2} (Fig. 4b), representing a record efficiency for perovskite LEDs. The brightness at the peak EQE is 18.4 W sr^{-1} m^{-2} (corresponding to a photon flux of 2.33×10^{20} m^{-2} s^{-1}), which is at least one order of magnitude higher than that of state-of-the-art organic LEDs (Extended Data Table 1). Owing to the low operation voltage and low EQE roll-off, the peak energy conversion efficiency (ECE, or wall–plug efficiency) reaches 18.6% at a current density of around 10 mA cm^{-2} and is maintained at 12.0% at a high current density of 100 mA cm^{-2}. Notably, the ECE at the current density of 100 mA cm^{-2} is much higher than that of the best-performing organic LEDs (Extended Data Table 1). An EQE histogram for 100 devices shows an average EQE of 19.2%, with a low relative standard deviation of 4% (Fig. 4c), indicating that the device performance is highly reproducible. The electroluminescence peak is located at 803 nm, consistent with the photoluminescence emission peak. The electroluminescence spectra do not vary at different viewing angles (Fig. 4d), and the angular emission intensity of our perovskite LEDs follows a Lambertian profile (Fig. 4e). These findings indicate that the randomly distributed submicrometre structure does not introduce a periodic grating effect. In addition, devices with a simple glass-epoxy encapsulation exhibit a half-lifetime (T_{50}, defined as the time taken for the EQE to drop to half of its initial value) of 20 h at the high constant current density of 100 mA cm^{-2} (Fig. 4f). This good stability can be attributed partially to the passivation effect of 5AVA, which can result in similar stability-enhancement effects in perovskite solar cells[26,27]. Moreover, we believe that the high ECE is beneficial to the perovskite stability, because less thermal energy is generated. We note that the T_{50} of 20 h at 100 mA cm^{-2} is also comparable with those of state-of-the-art near-infrared organic LEDs[30].

Considering that the PLQE of the perovskite layer is 70%, we suggest that the remarkable EQE of the electroluminescence is largely due to the enhanced light outcoupling efficiency of the spontaneously formed submicrometre structures shown in Figs. 1 and 2. To evaluate the enhancement factor of light outcoupling, we performed 3D finite-difference-time-domain (3D-FDTD) simulations (see Methods for details). The results show that a reference device with a continuous and flat emitting layer has an outcoupling efficiency of 21.8%. The outcoupling efficiency of a device with submicrometre structures reaches about 30% (Extended Data Fig. 5), which is consistent with the high EQE obtained in our devices. We also characterized our LED device under low temperatures, where the nonradiative recombination can be suppressed as the trap states are frozen out[24]. In this case, we can assume that the internal quantum efficiency of the LED device reaches nearly 100%, and the measured EQE is the outcoupling efficiency of the LED devices. Extended Data Fig. 6 shows that the device EQE increases with decreasing temperature and reaches 30% at 6 K, which is consistent with our 3D-FDTD simulation results.

We highlight that the device performance depends on several processing conditions, such as the ratio of 5AVA to FAI and PbI_2 (Extended Data Fig. 4) and the concentration of the precursor solution (Extended Data Figs. 7, 8). As discussed above, the PLQEs of the perovskite films can be improved by adding 5AVA into the precursor solutions. The formation of the submicrometre structure can also be affected by the ratio of 5AVA and the concentration of the precursor solution. After adding 5AVA, faceted perovskite platelets with submicrometre structure gradually form (Extended Data Fig. 4). Meanwhile, the PLQE and photoluminescence lifetime are greatly improved owing to a reduced defect density, leading to enhanced device efficiency (Extended Data Fig. 4). With more concentrated precursor solutions, there is a tendency to form larger crystallites as well as thicker and higher-surface-coverage films, which can greatly affect the leakage current, turn-on voltage, radiance and light outcoupling (Extended Data Fig. 7). When a solution of lower concentration is used, the devices show larger leakage current and lower EQEs. With thicker perovskite films, the devices become less conductive, which increases the turn-on voltage and reduces the brightness. Our current-processing parameters represent optimized conditions, which lead to perovskite films with a combination of high PLQE, low leakage current and preferred submicrometre structures.

We think that the PLQE and outcoupling efficiency could be enhanced further by using more effective additives to suppress the nonradiative recombination, and more optimized processing to control the submicrometre structure. Interestingly, conventional 3D perovskites are usually very sensitive to the fabrication process, but our device can maintain an average peak EQE of more than 18% with an annealing time of between 14 min and 20 min (Extended Data Fig. 2), showing that this solution-processing protocol has a relatively wide processing window.

We have also found that adding amino acids into the precursor solution is a general strategy for growing high-quality $FAPbI_3$ perovskites with submicrometre structure. Amino acids of different chain lengths—6-aminocaproic acid (6ACA) and 7-aminoheptanoic acid (7AHA)—have similar effects to 5AVA in achieving high-efficiency perovskite LEDs (Extended Data Fig. 9). After 6ACA or 7AHA is added to the precursor solution, the resulting films have morphologies and PLQEs analogous to those of films containing 5AVA. Without extensive optimization, the LEDs processed from precursor solutions with 6ACA or 7AHA also exhibit good peak EQEs of 18.2% and 17.3%, respectively.

It has been demonstrated that the EQEs of planar-type LEDs can be enhanced by improving the light outcoupling. However, traditionally this requires complex fabrication processes, and it is difficult to maintain a consistent emission spectrum at different viewing angles. Remarkably, these limitations can be avoided in perovskite LEDs by using the simple strategy described here, at little extra cost of fabrication. The resulting peak EQE of our perovskite LED approaches those of the best-performing organic LEDs. In contrast to LEDs based on group III–V semiconductors, in organic LEDs processed at low temperatures it is difficult to maintain high ECEs at high current densities, owing to their excitonic nature and low charge mobilities[1]. But low-temperature solution-processed perovskite LEDs demonstrate remarkably high ECEs at high current densities, suggesting the unique possibility of achieving large planar LEDs with high efficiency at high brightness.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0576-2.

Received: 11 February; Accepted: 22 August 2018;
Published online 10 October 2018.

1. Lee, J. et al. Deep blue phosphorescent organic light-emitting diodes with very high brightness and efficiency. *Nat. Mater.* **15**, 92–98 (2016).
2. Dai, X. et al. Solution-processed, high-performance light-emitting diodes based on quantum dots. *Nature* **515**, 96–99 (2014).
3. Waltereit, P. et al. Nitride semiconductors free of electrostatic fields for efficient white light-emitting diodes. *Nature* **406**, 865–868 (2000).
4. Lee, Y.-J. et al. A high-extraction-efficiency nanopatterned organic light-emitting diode. *Appl. Phys. Lett.* **82**, 3779–3781 (2003).
5. Bulović, V. et al. Weak microcavity effects in organic light-emitting devices. *Phys. Rev. B* **58**, 3730–3740 (1998).
6. Matterson, B. J. et al. Increased efficiency and controlled light output from a microstructured light-emitting diode. *Adv. Mater.* **13**, 123–127 (2001).
7. Ziebarth, J. M., Saafir, A. K., Fan, S. & McGehee, M. D. Extracting light from polymer light-emitting diodes using stamped bragg gratings. *Adv. Funct. Mater.* **14**, 451–456 (2004).
8. Sun, Y. & Forrest, S. R. Enhanced light out-coupling of organic light-emitting devices using embedded low-index grids. *Nat. Photon.* **2**, 483–487 (2008).
9. Koo, W. H. et al. Light extraction from organic light-emitting diodes enhanced by spontaneously formed buckles. *Nat. Photon.* **4**, 222–226 (2010).
10. Deschler, F. et al. High photoluminescence efficiency and optically pumped lasing in solution-processed mixed halide perovskite semiconductors. *J. Phys. Chem. Lett.* **5**, 1421–1426 (2014).
11. Tan, Z.-K. et al. Bright light-emitting diodes based on organometal halide perovskite. *Nat. Nanotechnol.* **9**, 687–692 (2014).
12. Wang, J. et al. Interfacial control toward efficient and low-voltage perovskite light-emitting diodes. *Adv. Mater.* **27**, 2311–2316 (2015).
13. Wang, N. et al. Perovskite light-emitting diodes based on solution-processed self-organized multiple quantum wells. *Nat. Photon.* **10**, 699–704 (2016).
14. Yuan, M. et al. Perovskite energy funnels for efficient light-emitting diodes. *Nat. Nanotechnol.* **11**, 872–877 (2016).
15. Yang, X. et al. Efficient green light-emitting diodes based on quasi-two-dimensional composition and phase engineered perovskite with surface passivation. *Nat. Commun.* **9**, 570; correction **9**, 1169 (2018).
16. Sun, Y. et al. The formation of perovskite multiple quantum well structures for high performance light-emitting diodes. *npj Flex. Electron.* **2**, 12 (2018).
17. Zou, W. et al. Minimising efficiency roll-off in high-brightness perovskite light-emitting diodes. *Nat. Commun.* **9**, 608 (2018).
18. Richter, J. M. et al. Enhancing photoluminescence yields in lead halide perovskites by photon recycling and light out-coupling. *Nat. Commun.* **7**, 13941 (2016).
19. Schmidt, L. C. et al. Nontemplate synthesis of $CH_3NH_3PbBr_3$ perovskite nanoparticles. *J. Am. Chem. Soc.* **136**, 850–853 (2014).
20. Sichert, J. A. et al. Quantum size effect in organometal halide perovskite nanoplatelets. *Nano Lett.* **15**, 6521–6527 (2015).
21. Lin, X. et al. Electrically-driven single-photon sources based on colloidal quantum dots with near-optimal antibunching at room temperature. *Nat. Commun.* **8**, 1132 (2017).
22. Fu, Y. et al. Stabilization of the metastable lead iodide perovskite phase via surface functionalization. *Nano Lett.* **17**, 4405–4414 (2017).
23. Pang, S. et al. $NH_2CH{=}NH_2PbI_3$: an alternative organolead iodide perovskite sensitizer for mesoscopic solar cells. *Chem. Mater.* **26**, 1485–1491 (2014).
24. Stranks, S. D. et al. Recombination kinetics in organic-inorganic perovskites: excitons, free charge, and subgap states. *Phys. Rev. Appl.* **2**, 034007 (2014).
25. Gélvez-Rueda, M. C., Renaud, N. & Grozema, F. C. Temperature dependent charge carrier dynamics in formamidinium lead iodide perovskite. *J. Phys. Chem. C* **121**, 23392–23397 (2017).
26. Mei, A. et al. A hole-conductor-free, fully printable mesoscopic perovskite solar cell with high stability. *Science* **345**, 295–298 (2014).
27. Zhang, T. et al. In situ fabrication of highly luminescent bifunctional amino acid crosslinked 2D/3D $NH_3C_4H_9COO(CH_3NH_3PbBr_3)_n$ perovskite films. *Adv. Funct. Mater.* **27**, 1603568 (2017).
28. de Quilettes, D. W. et al. Impact of microstructure on local carrier lifetime in perovskite solar cells. *Science* **348**, 683–686 (2015).
29. Lee, S. et al. Amine-based passivating materials for enhanced optical properties and performance of organic-inorganic perovskites in light-emitting diodes. *J. Phys. Chem. Lett.* **8**, 1784–1792 (2017).
30. Sassi, M. et al. Near-infrared roll-off-free electroluminescence from highly stable diketopyrrolopyrrole light emitting diodes. *Sci. Rep.* **6**, 34096 (2016).

Acknowledgements This work is supported financially by the Joint Research Program between China and the European Union (2016YFE0112000); the Major Research Plan of the National Natural Science Foundation of China (91733302); the National Basic Research Program of China-Fundamental Studies of Perovskite Solar Cells (2015CB932200); the Natural Science Foundation of Jiangsu Province, China (BK20150043, BK20150064, BK20180085); the National Natural Science Foundation of China (11474164, 11474249, 61634001); the National Science Fund for Distinguished Young Scholars (61725502, 61725503); and the Synergetic Innovation Center for Organic Electronics and Information Displays. We thank D. Di and B. Zhao for cross-checking the LED measurement system and Y. Zhao for helpful discussions. We thank M. Winton and N. Greenham for proof reading.

Reviewer information *Nature* thanks A. Urban and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.W. had the idea for and designed the experiments. J.W. and W.H. supervised the work. Y.C. carried out device fabrication and characterizations, with the assistance of Y.M., H.Cao and Y.K.; Y.C., W.Z., M.X., Y.Wang, Z.F., D.K., Q.P., M.Y. and Y.H. conducted the optical measurements. Y.Wei and H.L. carried out AFM measurements. H.Chen, G.L. and Y.J. carried out FTIR characterizations. H.T. carried out high-resolution TEM and STEM characterizations with the assistance of Y.C., Y.Wei and K.D.; J.G. carried out optical simulations of the device with the assistance of K.P.; D.D. supervised the optical simulation. J.W., N.W. and Y.C. analysed the data. N.W. wrote the first draft of the manuscript and J.W. and W.H. provided major revisions. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0576-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.W. or W.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Device fabrication. ZnO nanocrystals were spin-coated onto ITO-coated glass substrates, forming the electron-transport layer. Then an ultrathin PEIE layer was fabricated onto the ZnO layer to decrease the work function and improve the wetting property of ZnO[12,13,31]. A precursor perovskite solution was prepared by dissolving 5AVA, FAI and PbI_2 with different molar ratios and concentrations in dimethylformamide (DMF), and stirring at 60 °C for 2 h in a nitrogen-filled glovebox. A perovskite precursor containing 6ACA (or 7AHA) was prepared by dissolving 6ACA (or 7AHA), FAI and PbI_2 with a molar ratio of 0.7/2.4/1 in DMF. Next, the perovskite films were prepared by spin-coating the precursor solution onto the PEIE-treated ZnO films and annealing them at 100 °C. TFB was spin-coated onto the perovskite film, forming the hole-transport layer. Finally, an MoO_x/Au electrode was thermally evaporated through a shadow mask, defining the device area of 3 mm^2.

Device characterization. We tested the perovskite LEDs on top of an integration sphere at room temperature in a nitrogen-filled glovebox, in which only forward light emission can be collected[2,32]. We measured the devices from zero bias to forward bias at a rate of 0.05 V s^{-1}, and recorded the data with the first scan without pre-bias. We carried out low-temperature characterizations of the LEDs in a cryostat (Oxford Instruments NanoScience, OptistatAC-V12W). The stability of the devices was measured in air with simple glass-epoxy encapsulation. We measured the angular dependence of emission intensity and spectra using a Thorlabs PDA100A detector and QE65 Pro spectrometer, respectively. Three sets of perovskite LEDs from the same batch were cross-checked at Nanjing Tech University, the University of Cambridge (Optoelectronics Group) and Zhejiang University (Y. Jin group), and the results are consistent (Extended Data Table 2).

Film characterizations. We collected STEM images of cross-sectional devices on a FEI Titan G2 80-200 ChemiSTEM operated at 200 keV. We used the FEI Talos analytical FEG scanning transmission electron microscope, which includes the Super-X energy-dispersive X-ray spectroscopy (EDS) system with four silicon drift detectors for superior sensitivity and mapping capabilities. We acquired images every 2° in the range −150° to +150° to avoid a missing-wedge effect, with an 80-keV high voltage and 200-pA incident beam current to avoid beam damage. Visualization and reconstruction was done using FEI Inspect 3D and Avizo software. We obtained SEM images of perovskite films with a JEOL5 JSM-7800F SEM. The surface morphology of different layers was collected by AFM (Bruker, Dimension ICON).

We used an ultraviolet/visible spectrophotometer with an integrating sphere (PerkinElmer, Lambda 950) to measure the absorbance spectra. Photoluminescence spectra were measured using a QE65 Pro spectrometer and a 445-nm continuous-wave laser as an excitation source. Time-resolved photoluminescence measurements were performed with an Edinburgh Instruments spectrometer (FLS980), having excited the perovskite films with a 633-nm pulsed laser of various intensities. We measured the PLQE of perovskite films by combining a continuous-wave laser, optical fibre, spectrometer and integrating sphere[33].

We collected XRD data using a Bruker D8 Advance. Grazing-angle FTIR spectra were performed with a Thermo Fisher IS50 equipped with a Smart SAGA reflectance accessory. The samples were prepared on Au-coated quartz substrates.

Simulations. We analysed the spatial frequency spectrum of the randomly distributed perovskite map of a real device by using a two-dimensional fast Fourier transform (FFT). The range P_{range} is given as P_1–P_2, corresponding to the major spatial frequency components. 3D-FDTD (Lumerical Solutions) simulations were carried out for the regular periodic devices with a cycle P ($l_e = P/2$, where l_e is the length of the perovskite platelets) to estimate the outcoupling efficiency, $\eta_{FDTD}(P)$. Then the average value of the calculated efficiency $\eta_{FDTD}(P_1)$, $\eta_{FDTD}(P_1 + 1\ \text{nm})$, … $\eta_{FDTD}(P_2)$ was used to evaluate the outcoupling efficiency of the real device (with a wideband spatial frequency spectrum). The error corresponds to the standard deviation.

The SEM data were imported and a picture with $N \times N$ pixels was obtained, which was then discretized as the function:

$$f(x,y)=\begin{cases}1\ (\text{EML area})\\ 0\ (\text{else area})\end{cases}$$

where x, $y = a$, $2a$…, Na, and a is the pixel size. The spatial frequency spectrum:

$$F(U_x,U_y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\mathrm{e}^{-j2\pi(U_x x+U_y y)}\mathrm{d}x\mathrm{d}y$$

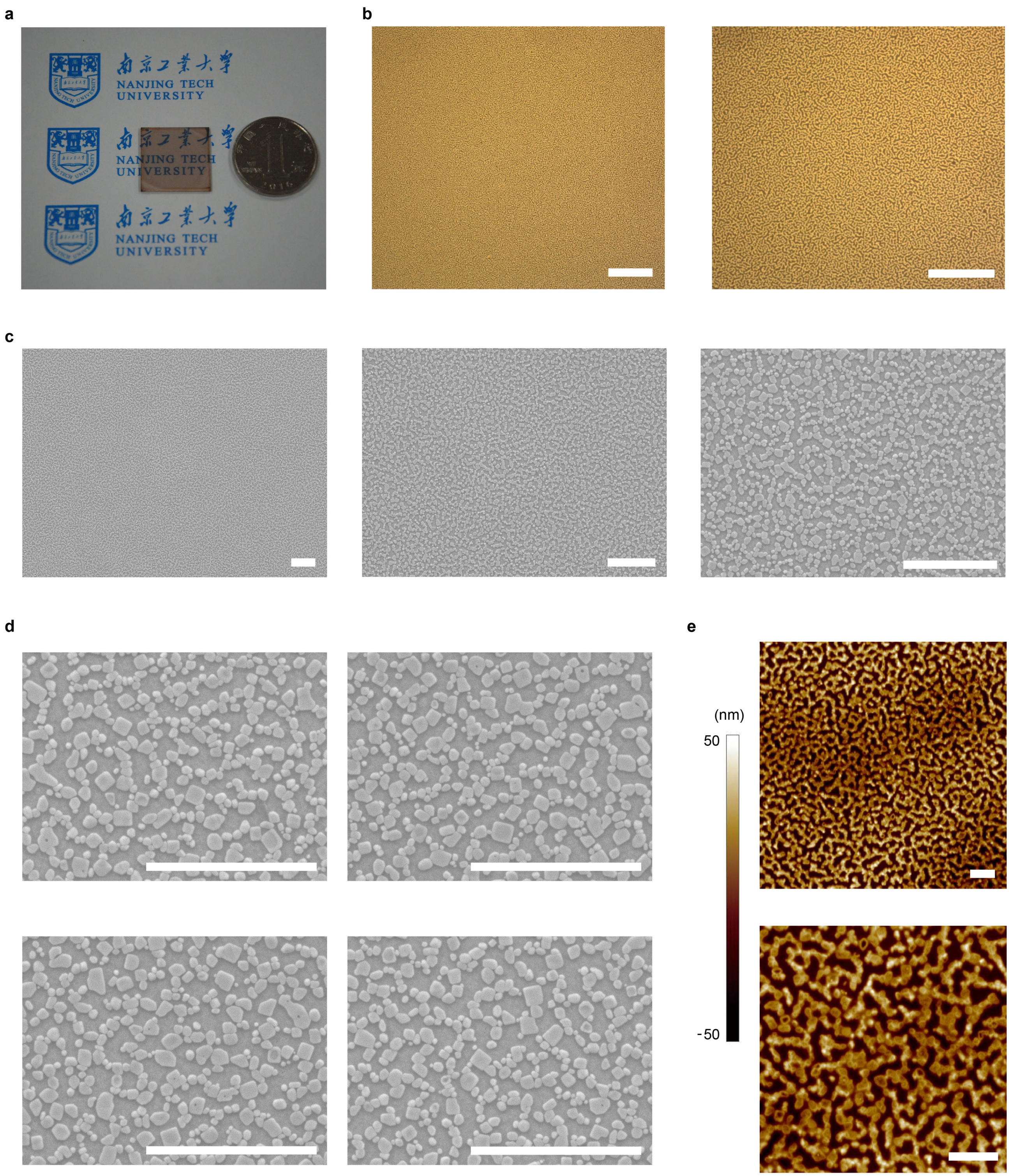
was obtained by using FFT (Extended Data Fig. 5c). For example, we can obtain a discretized picture with 1,001 × 1,001 pixels and $a = 3.57$ nm (Extended Data Fig. 5b). The major peaks of $|F(U_x, U_y)|$ are found in the range $r_1{\cdot}(2\pi/a) \le |U_x|$, $|U_y| \le r_2{\cdot}(2\pi/a)$ ($r_1 = 3.52 \times 10^{-3}$ and $r_2 = 9.83 \times 10^{-3}$, as labelled with white squares in Extended Data Fig. 5c). Then P_{range} can be established as 363–1,015 nm ($P_1 = a/r_2$, $P_2 = a/r_1$), resulting in an outcoupling efficiency of 28.7% ± 2.6% when the convex height, h, is 30 nm.

For the 3D-FDTD simulations, we used a window of 7 μm × 7 μm × 0.45 μm, and applied localized refined meshes (Δx, Δy and Δz). We chose $\Delta x = \Delta y = \Delta z = 2$ nm in the dipole source area, and $\Delta z = 1$ nm in the convex structures. In order to simulate the incoherent isotropic light-generation process, we did simulations thrice by choosing different source polarizations for the dipole source. The near field at the glass–ITO interface was then used to calculate the outcoupled far field in air by using Lumerical's far field analysis group, in which the air–glass interface was taken into consideration. We measured the refractive indices of Au, MoO_x, TFB, perovskite and ZnO with an ellipsometer (KLA Tencor, P-7), and took the refractive indices of ITO and glass from the literature[34,35]. To obtain the optical constants of perovskite in the simulation, we prepared a continuous $FAPbI_3$ film for the ellipsometer measurement, and fitted the measurement result using the dispersion laws of the Tauc–Lorentz model, the Gauss model and two sets of Lorentz oscillators. We ignored the imaginary part of the refractive index of the perovskite in the simulation. We carried out the simulation at a wavelength of 800 nm. For each periodic device, we used 3 × 3 uniform distributed sources in one perovskite cubic in separate simulations. Dividing the average outcoupling power by the average source power obtained from all the simulations (varying the locations and polarizations of sources), we obtained the theoretical outcoupling efficiency of the device.

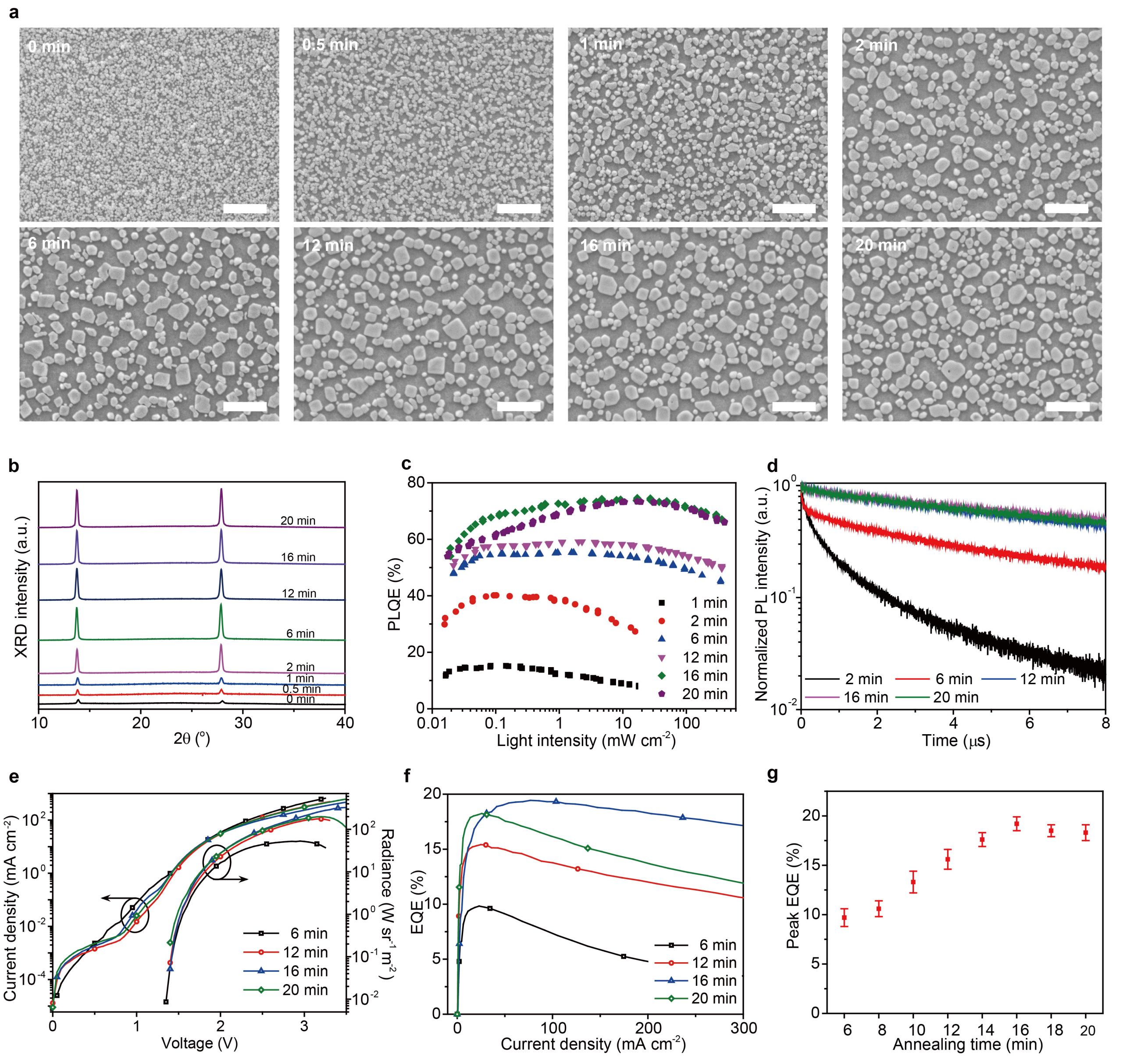
Data availability

The data that support the finding of this study are available from the corresponding author upon reasonable request.

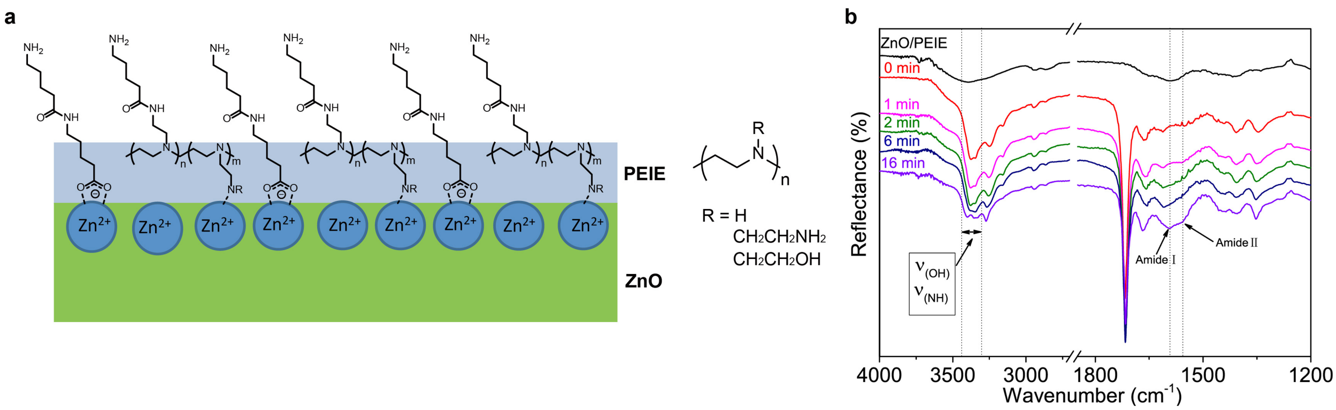
31. Zhou, Y. et al. A universal method to produce low-work function electrodes for organic electronics. *Science* **336**, 327–332 (2012).
32. Forrest, S. R., Bradley, D. D. C. & Thompson, M. E. Measuring the efficiency of organic light-emitting devices. *Adv. Mater.* **15**, 1043–1048 (2003).
33. de Mello, J. C., Wittmann, H. F. & Friend, R. H. An improved experimental determination of external photoluminescence quantum efficiency. *Adv. Mater.* **9**, 230–232 (1997).
34. Rubin, M. Optical properties of soda lime silica glasses. *Sol. Energy Mater.* **12**, 275–288 (1985).
35. König, T. A. F. et al. Electrically tunable plasmonic behavior of nanocube–polymer nanomaterials induced by a redox-active electrochromic polymer. *ACS Nano* **8**, 6182–6192 (2014).
36. Ogomi, Y. et al. All-solid perovskite solar cells with HOCO-R-$NH_3^+I^-$ anchor-group inserted between porous titania and perovskite. *J. Phys. Chem. C* **118**, 16651–16659 (2014).
37. Tamaddon, F., Aboee, F. & Nasiri, A. ZnO nanofluid as a structure base catalyst for chemoselective amidation of aliphatic carboxylic acids. *Catal. Commun.* **16**, 194–197 (2011).
38. Ly, K. T. et al. Near-infrared organic light-emitting diodes with very high external quantum efficiency and radiance. *Nat. Photon.* **11**, 63–68 (2017).
39. Kim, K.-H. et al. Phosphorescent dye-based supramolecules for high-efficiency organic light-emitting diodes. *Nat. Commun.* **5**, 4769 (2014).
40. Helander, M. G. et al. Chlorinated indium tin oxide electrodes with high work function for organic device compatibility. *Science* **332**, 944–947 (2011).
41. Lin, T.-A. et al. Sky-blue organic light emitting diode with 37% external quantum efficiency using thermally activated delayed fluorescence from spiroacridine-triazine hybrid. *Adv. Mater.* **28**, 6976–6983 (2016).



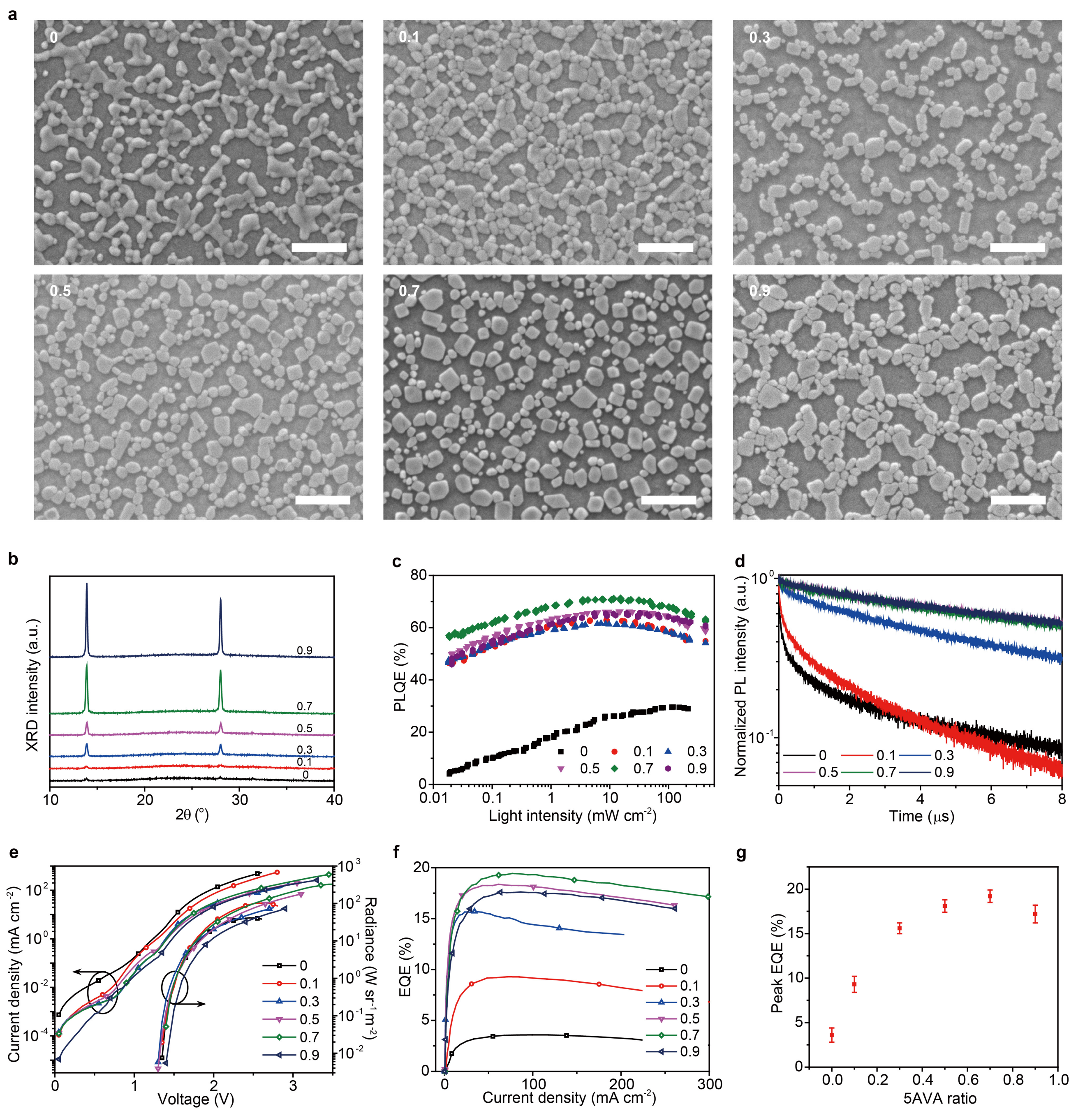
Extended Data Fig. 1 | Images of perovskite films. a, Photograph of a perovskite film on a 2 cm × 2 cm glass/ZnO-PEIE substrate, alongside a coin. The perovskite film is shiny and uniform. **b**, Optical microscope images with different magnifications. The scale bars represent 30 μm. **c**, SEM images with different magnifications. The scale bars represent 3 μm. **d**, High-magnification SEM images of randomly selected regions. The scale bars represent 3 μm. **e**, AFM images with different magnifications. The scale bars represent 2 μm. The images show the submicrometre-scale structure of the perovskite film.



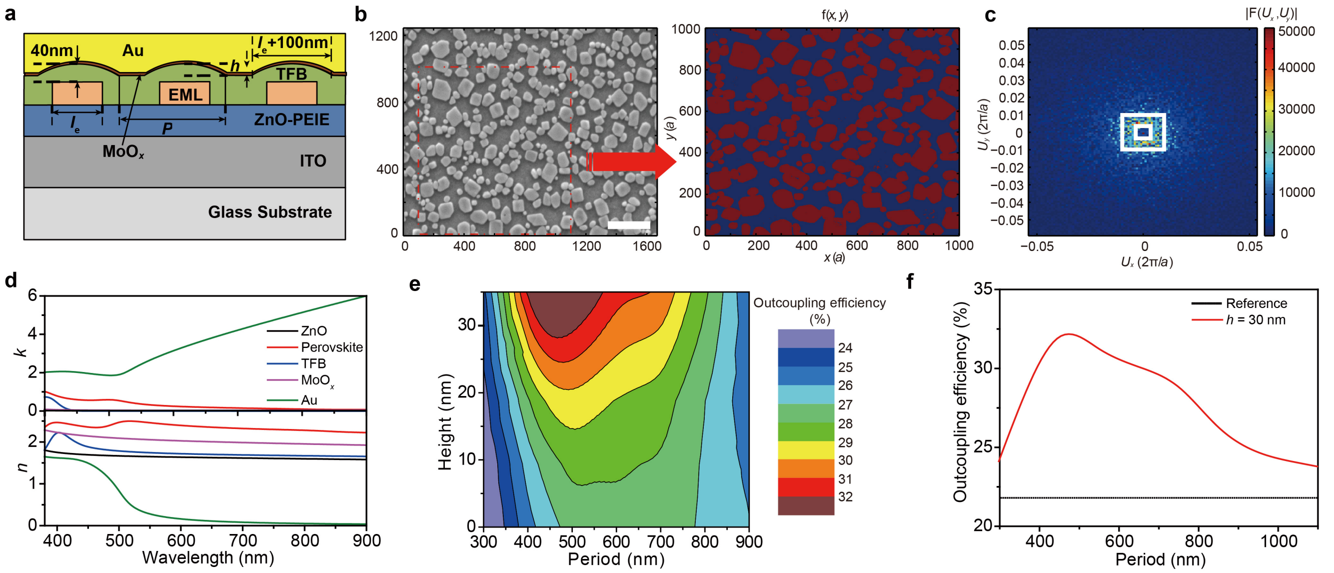
Extended Data Fig. 2 | Characterization of our perovskite films and perovskite LEDs fabricated with different annealing times. a, SEM images of perovskite films. The scale bars represent 1 μm. The images show that as the annealing time increases, the crystallites grow from small particles and become larger and more faceted. When the annealing time is more than 6 min, similar submicrometre-scale structures form. **b**, XRD spectra. Crystallinity is enhanced as the annealing time increases, in agreement with the SEM images. **c**, Excitation-intensity-dependent PLQEs. Trap densities gradually decrease as the annealing time increases, resulting in PLQEs of more than 60% when the annealing time is between 16 and 20 min. **d**, Time-resolved photoluminescence (PL) decay transients (at a carrier density of 1.0×10^{13} cm^{-3}). The films show longer PL decay lifetimes at longer annealing times. **e**, Dependence of current density and radiance on the driving voltage. The circles denote a bunch of curves; the arrows show the *y* axis to which a given bunch belongs. **f**, EQE versus current density. **g**, Peak EQE versus annealing time. An average peak EQE of more than 18% can be maintained with annealing times of between 14 and 20 min. Error bars correspond to the standard deviation.



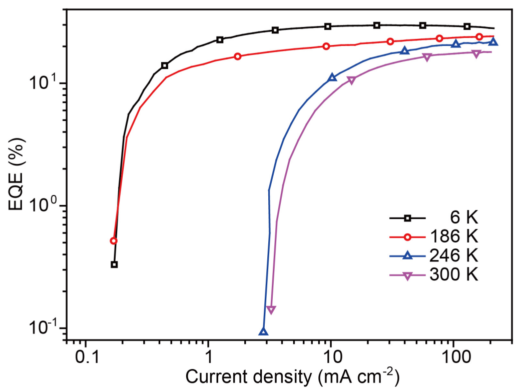
Extended Data Fig. 3 | Formation of the organic layer surrounding the submicrometre structures. a, Dehydration reaction of 5AVA on top of the ZnO-PEIE surface[36,37]. **b**, Grazing-angle reflectance FTIR spectra of perovskite films at various annealing times. As the annealing time increases, the peak at 3,400–3,300 cm^{-1} (ν_{O-H} in 5AVA) decreases; simultaneously, peaks appear at 1,620 cm^{-1} and 1,550 cm^{-1}, which can be assigned as amide I band ($\nu_{C=O}$) and amide II band (δ_{N-H}), respectively. These spectra indicate dehydration reactions of 5AVA, leading to the formation of an organic layer during annealing.



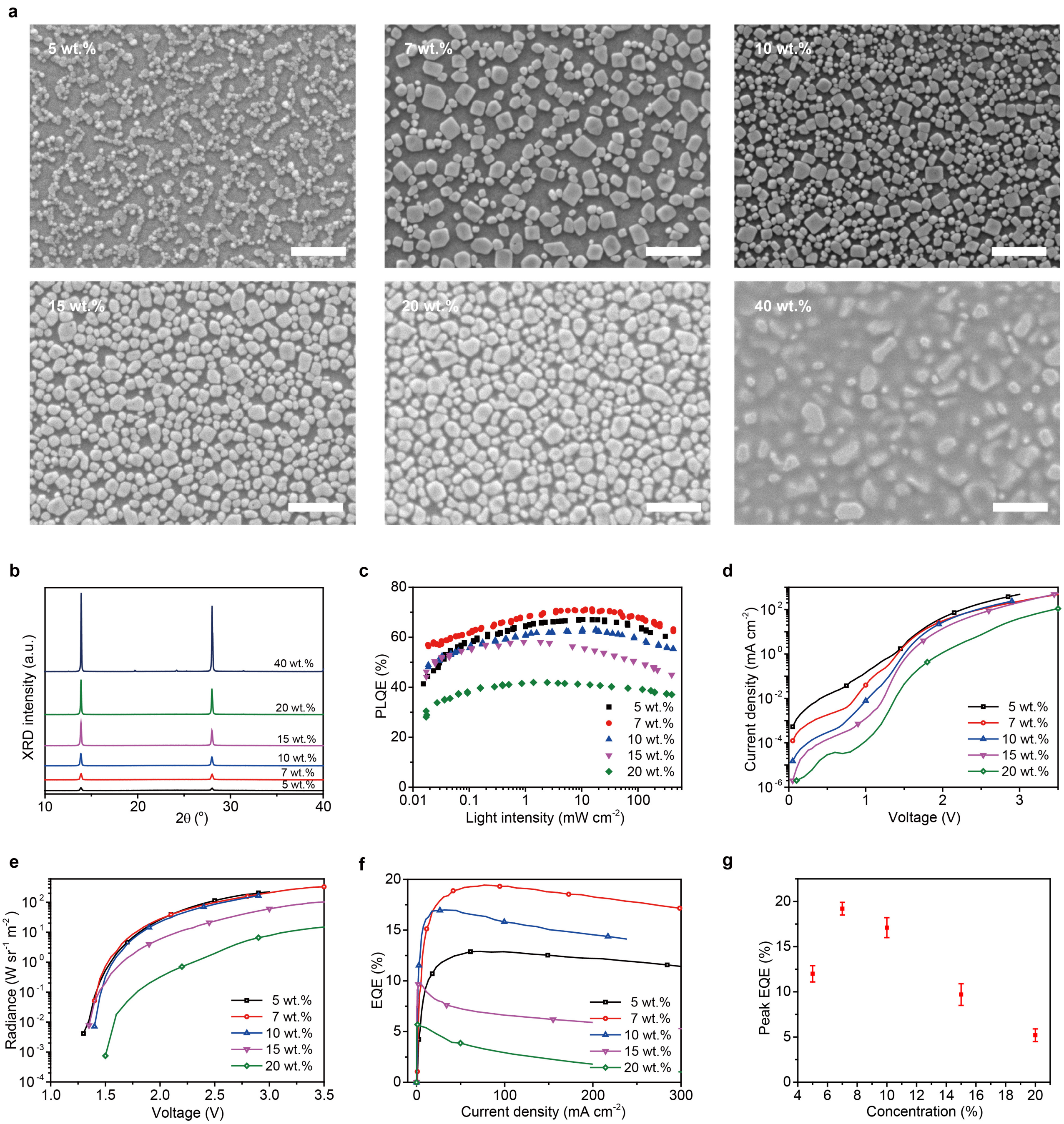
Extended Data Fig. 4 | Characterizations of perovskite films and LEDs with various 5AVA amounts. The ratio of 5AVA to FAI to PbI_2 is x/2.4/1, where x varies from 0 to 0.9. **a**, SEM images. The scale bars represent 1 μm. The value of x is given in the top left corner of each image. The reference $FAPbI_3$ perovskite film without 5AVA has low film coverage. Without 5AVA, the perovskites form discrete clusters with random shapes. After adding 5AVA, faceted perovskites with submicrometre structures gradually form. **b**, XRD spectra. The perovskite films show improved crystallinity with the addition of 5AVA. **c**, Excitation-intensity-dependent PLQE. After adding 5AVA, PLQEs were greatly enhanced, indicating reduced trap densities. **d**, Time-resolved PL decay transients (carrier density 1.0×10^{13} cm^{-3}). There is a fast PL decay channel for the perovskite without 5AVA, indicating a high level of trap densities. This fast PL decay channel gradually disappears after adding 5AVA. **e**, Dependence of current density and radiance on the driving voltage. After adding 5AVA, the leakage current is reduced. **f**, EQE plotted against current density. **g**, Peak EQE plotted against 5AVA ratio. Error bars correspond to the standard deviation. After adding 5AVA, the peak EQE increases owing to reduced leakage current and enhanced PLQE. When the 5AVA ratio is increased to 0.9, the EQE decreases, owing to the inferior outcoupling efficiency that results from the more dispersed structural pattern.



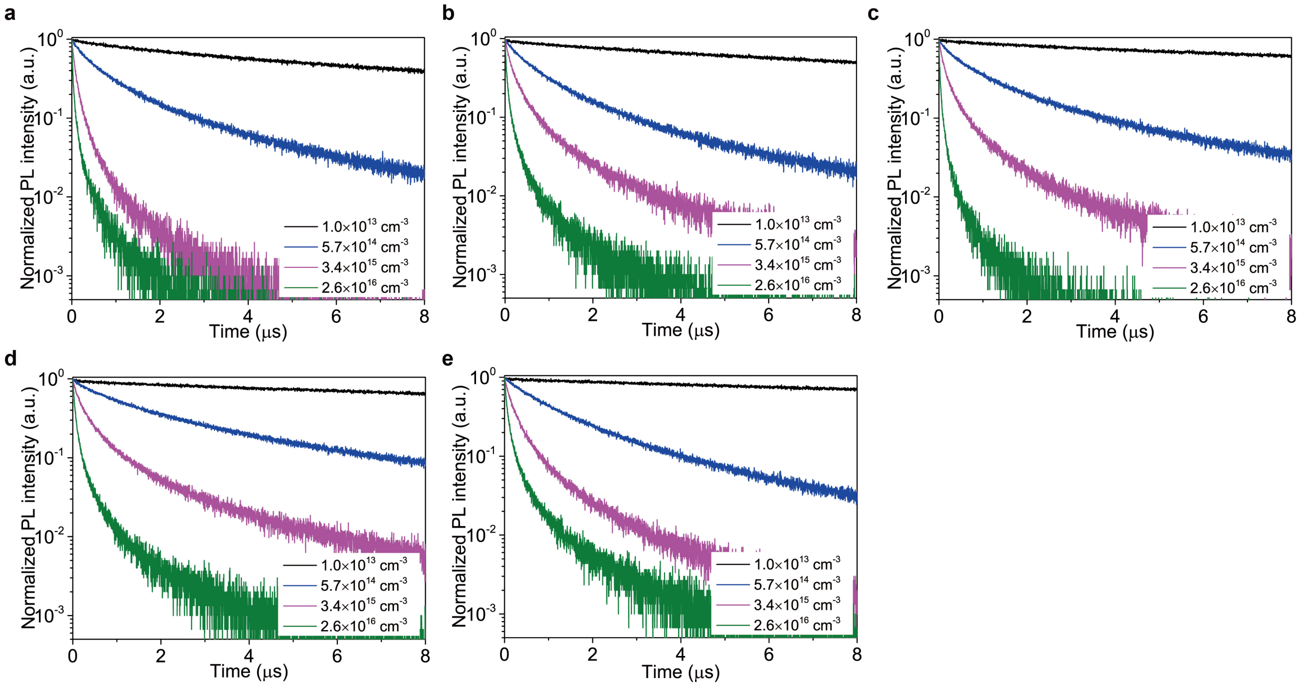
Extended Data Fig. 5 | Simulation of outcoupling efficiency. a, Device structure. A typical reference device consists of a metal layer (Au), a 7-nm-thick MoO_3 layer, a 40-nm-thick TFB layer, a 50-nm-thick emitting layer (EML), a 30-nm-thick layer of ZnO-PEIE, a 160-nm-thick ITO layer and a semi-infinite glass substrate. In our new device, the EML is replaced by a layer of perovskite squares distributed with a period P and a duty cycle l_e/P (where l_e is the length of the perovskite platelets, and $l_e/P = 50\%$). The height of the convex structure of TFB is denoted as h and the diameter is set to $l_e + 100$ nm. **b**, Discretized map of the perovskite layer. The scale bar represents 1 μm. x and y are the pixel numbers in units of pixel length a. $f(x,y)$ is the discrete function. **c**, Module of spatial frequency spectrum. U_x and U_y are the spatial frequencies. **d**, Refractive indices of different layers in our perovskite LEDs. Optical constants (n, k) of the multilayers were determined using an ellipsometer. Here the optical constants of perovskite are from a continuous $FAPbI_3$ film, which are used in the simulation. **e**, EQE calculated as the period P and the convex height h. **f**, Calculated outcoupling efficiency as a function of period P with convex height $h = 30$ nm. The reference is a device made from continuous perovskite film. The simulation shows that the outcoupling efficiency can be more than 25% over a wide range of periods from 310 nm to 900 nm.



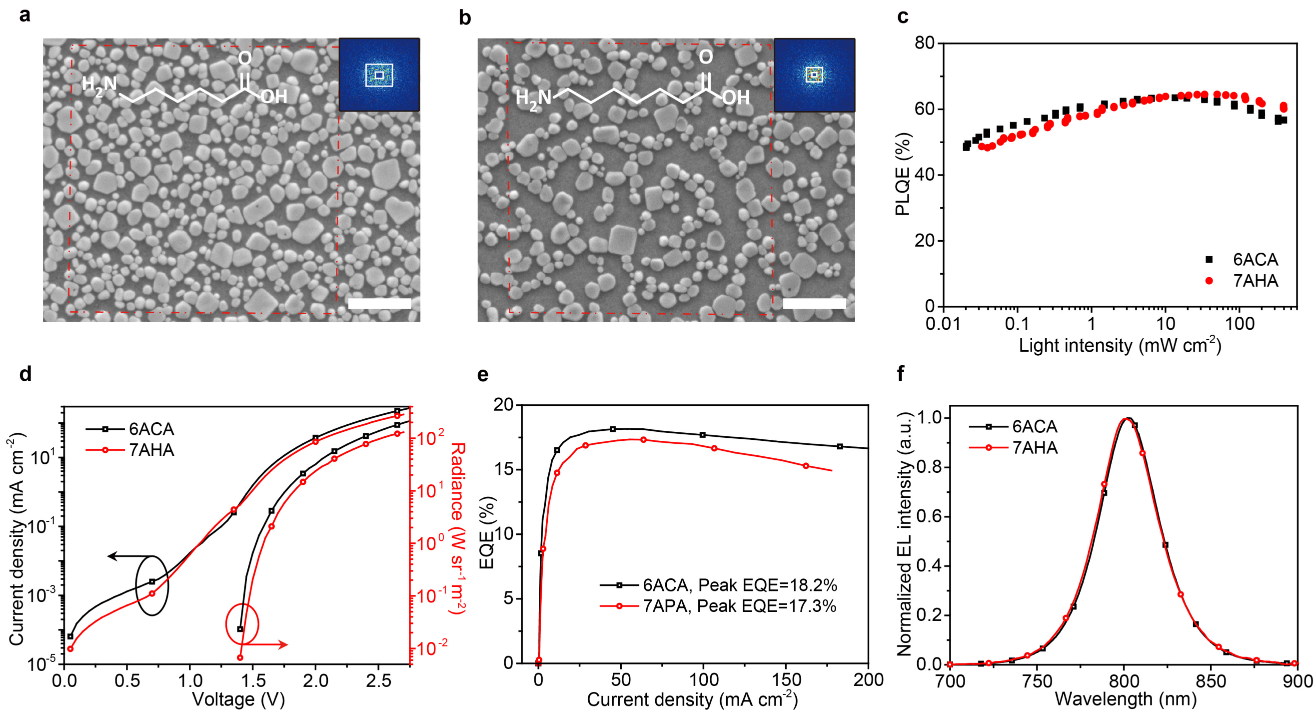
Extended Data Fig. 6 | EQE versus current density for our perovskite LED device at different temperatures. Measuring the device EQE at low temperatures minimizes nonradiative recombination so that the EQE reaches a value of 30% at 6 K.



Extended Data Fig. 7 | Characterization of our perovskite films and LEDs at different precursor concentrations. a, SEM images. The scale bars represent 1 μm. As the precursor concentration increases (shown in the top left corner of each image), the size of the crystallites increases and the crystallites become more tightly packed. **b**, XRD spectra. **c**, Excitation-intensity-dependent PLQEs. The 7 wt.% film shows the highest PLQEs. **d**, Dependence of current density on driving voltage. The leakage currents of the devices decrease as the precursor solution becomes more concentrated. **e**, Dependence of radiance on the driving voltage. As the precursor concentration increases, the turn-on voltage increases and the radiance decreases, probably owing to the poor charge transport of thicker film. **f**, EQE plotted against current density. **g**, Peak EQE versus precursor concentration. Error bars correspond to the standard deviation. When the concentration exceeds 10 wt.% the EQE decreases, probably because of a reduced outcoupling-enhancement effect and poor charge transport.



Extended Data Fig. 8 | Time-resolved photoluminescence decay transients of perovskites with different precursor concentrations. **a**, 5 wt.%. **b**, 7 wt.%. **c**, 10 wt.%. **d**, 15 wt.%. **e**, 20 wt.%. Charge carrier densities vary as indicated (black, blue, purple and green traces).



Extended Data Fig. 9 | Optoelectronic characteristics of perovskite LED devices fabricated with different amino acids in the precursor solution. **a**, SEM image of submicrometre-structured perovskites fabricated with 6ACA (chemical structure shown in white). The scale bar represents 1 μm. Inset, FFT pattern in a randomly selected region. The *P* range of 6ACA is 265–901 nm, yielding a calculated outcoupling efficiency of 28.9% ± 2.5%. **b**, SEM image of submicrometre-structured perovskites fabricated with 7AHA (chemical structure shown in white). The scale bar represents 1 μm. Inset, FFT pattern in a randomly selected region. The *P* range of 7AHA is 432–1,430 nm, yielding a calculated outcoupling efficiency of 26.4% ± 3.3%. **c**, Excitation-intensity-dependent PLQE. The perovskite films with 6ACA and 7AHA have similar PLQEs. **d**, Dependence of current density and radiance on the driving voltage. **e**, EQE versus current density. The 6ACA- and 7AHA-based devices reach peak EQEs of 18.2% and 17.3%, respectively. Given that the perovskite films based on 6ACA and 7AHA have similar PLQEs, the EQEs must be affected mainly by the different outcoupling efficiencies that result from the different periodicities of the submicrometre-scale structures. **f**, Electroluminescence spectra.

Extended Data Table 1 | Comparison of our device with reported high-performance organic LEDs

Device*	NIR		Red	Green	Blue
	Our LED	OLED[38]	OLED[39]	OLED[40]	OLED[41]
Peak EQE (%)	20.7	24	35.6	29.2	36.7
Current density (mA cm^{-2}) @ peak EQE	18	0.4†	0.7†	1†	0.1†
Photon flux ($\times 10^{20}$ m^{-2}s^{-1}) @ peak EQE	2.33	0.06‡	0.16‡	0.18‡	0.02‡
EQE (%) @ 100 mA cm^{-2}	19	-	-	19†	-
ECE (%) @ 100 mA cm^{-2}	12	-	-	5.6‡	-
EL peak (nm)	803	740	610	525	490†

Our perovskite LED device is compared with the organic LEDs (OLEDs) reported in refs [38–41].
*The table includes organic LEDs with horizontally orientated emitting dipoles, but without external optical outcoupling schemes (an example of an optical outcoupling scheme being a microlens with a glass substrate).
†These data were obtained from the figures in refs [38–41].
‡These data were estimated from the EQE, current density, voltage or photon energy at the electroluminescence peak.

Extended Data Table 2 | Comparison of devices measured in different laboratories

Device	Group	Peak EQE (%)†
PeLEDs*	Nanjing Tech University	15.2±0.3
	Cambridge University	15.3±0.5
	Zhejiang University	15.1±0.3

*The perovskite LEDs (PeLEDs) were fabricated in the same batch at Nanjing Tech University. Each set of devices without encapsulations was simultaneously measured in air until the Optoelectronics Group, Cambridge, received its samples. As the transfer time from Nanjing to Cambridge is about eight days, the devices suffer from modest degradations.
†The average peak EQE and standard deviation are from ten devices.

https://doi.org/10.1038/s41586-018-0574-4

Solution-processable 2D semiconductors for high-performance large-area electronics

Zhaoyang Lin[1], Yuan Liu[2,3], Udayabagya Halim[1], Mengning Ding[2], Yuanyue Liu[4], Yiliu Wang[1], Chuancheng Jia[2], Peng Chen[3], Xidong Duan[3], Chen Wang[2], Frank Song[1], Mufan Li[1], Chengzhang Wan[1], Yu Huang[2,5]* & Xiangfeng Duan[1,5]*

Two-dimensional (2D) materials, consisting of atomically thin crystal layers bound by the van der Waals force, have attracted much interest because of their potential in diverse technologies, including electronics, optoelectronics and catalysis[1–10]. In particular, solution-processable 2D semiconductor (such as MoS_2) nanosheets are attractive building blocks for large-area thin-film electronics. In contrast to conventional zero- and one-dimensional nanostructures (quantum dots and nanowires, respectively), which are typically plagued by surface dangling bonds and associated trapping states, 2D nanosheets have dangling-bond-free surfaces. Thin films created by stacking multiple nanosheets have atomically clean van der Waals interfaces and thus promise excellent charge transport[11–15]. However, preparing high-quality solution-processable 2D semiconductor nanosheets remains a challenge. For example, MoS_2 nanosheets and thin films produced using lithium intercalation and exfoliation are plagued by the presence of the metallic 1T phase and poor electrical performance (mobilities of about 0.3 square centimetres per volt per second and on/off ratios of less than 10)[2,12], and materials produced by liquid exfoliation exhibit an intrinsically broad thickness distribution, which leads to poor film quality and unsatisfactory thin-film electrical performance (mobilities of about 0.4 square centimetres per volt per second and on/off ratios of about 100)[14,16,17]. Here we report a general approach to preparing highly uniform, solution-processable, phase-pure semiconducting nanosheets, which involves the electrochemical intercalation of quaternary ammonium molecules (such as tetraheptylammonium bromide) into 2D crystals, followed by a mild sonication and exfoliation process. By precisely controlling the intercalation chemistry, we obtained phase-pure, semiconducting 2H-MoS_2 nanosheets with a narrow thickness distribution. These nanosheets were then further processed into high-performance thin-film transistors, with room-temperature mobilities of about 10 square centimetres per volt per second and on/off ratios of 10^6 that greatly exceed those obtained for previous solution-processed MoS_2 thin-film transistors. The scalable fabrication of large-area arrays of thin-film transistors enabled the construction of functional logic gates and computational circuits, including an inverter, NAND, NOR, AND and XOR gates, and a logic half-adder. We also applied our approach to other 2D materials, including WSe_2, Bi_2Se_3, $NbSe_2$, In_2Se_3, Sb_2Te_3 and black phosphorus, demonstrating its potential for generating versatile solution-processable 2D materials.

Lithium intercalation is a common approach to exfoliating layered crystals. During lithium intercalation, the insertion of each Li^+ ion involves the injection of one electron into the host crystal. The intercalation of a large number of Li^+ ions (one per formula unit in $LiMoS_2$) therefore leads to massive electron injection into the MoS_2 crystal, which induces an undesired phase transition from the semiconducting 2H phase to the metallic 1T phase[2,5,11]. Previous theoretical studies suggest that this phase transition occurs only when the electron injection exceeds a certain threshold (0.29 electrons per MoS_2 formula unit)[18,19]. A plausible way of reducing electron injection into the host 2D crystal and thus preventing the undesired phase transition is to replace the small Li^+ ions (diameter $d \approx 2$ Å) with larger cations, such as quaternary ammonium ions. The large size of quaternary ammonium molecules such as tetraheptylammonium bromide (THAB; $d \approx 20$ Å) naturally limits the number of molecules that can fit into the host crystal and thus the number of electrons injected.

We inserted THAB molecules into the MoS_2 crystal in an electrochemical cell[20] (see Methods, Extended Data Fig. 1). Driven by the

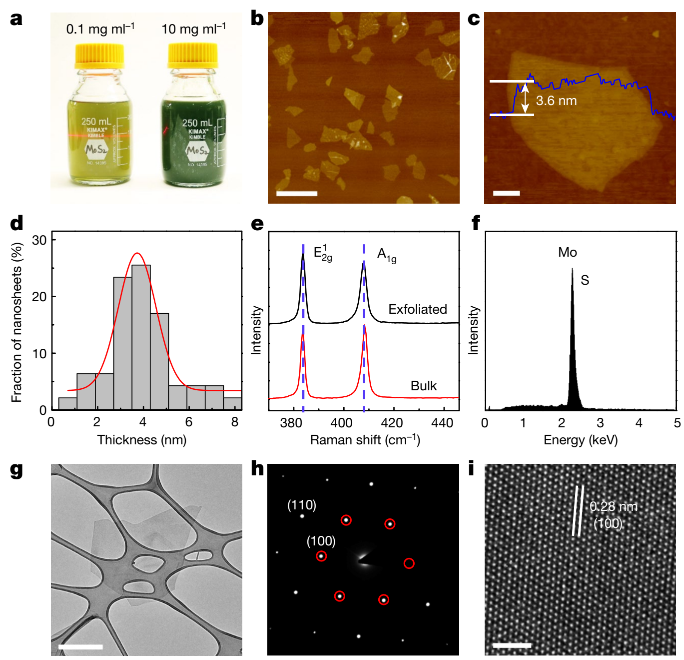


Fig. 1 | Structural characterizations of exfoliated MoS_2 nanosheets. **a**, Photograph of exfoliated MoS_2 nanosheets dispersed in isopropanol (ink solution) at different concentrations. The Tyndall effect was observed in the 0.1 mg ml^{-1} solution. **b**, AFM image of multiple MoS_2 nanosheets with a narrow thickness distribution. Scale bar, 2 μm. **c**, AFM image of an individual MoS_2 nanosheet. The blue line represents the height profile across the nanosheet, which indicates a thickness of 3.6 nm. Scale bar, 200 nm. **d**, The thickness distribution of exfoliated MoS_2 nanosheets as measured by AFM (bars) and a Gaussian fit (red line). The mean thickness is 3.8 nm and the standard deviation is 0.9 nm. **e**, Raman spectroscopy analysis of the exfoliated nanosheets (top, black), and of the bulk crystal for comparison (bottom, red). The vertical dashed lines indicate the positions of the two Raman peaks. **f**, EDS spectrum obtained from exfoliated MoS_2 nanosheets. **g–i**, A typical TEM image (**g**), selected-area electron diffraction image (**h**) and high-resolution TEM image (**i**) of a single MoS_2 nanosheet. The red circles in **h** outline the six-fold-symmetric diffraction spots. The labelled plane in **i** is indexed to be the (100) plane, with 0.28 nm lattice spacing. Scale bars, 500 nm (**g**) and 2 nm (**i**).

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA. [2]Department of Materials Science and Engineering, University of California, Los Angeles, CA, USA. [3]State Key Laboratory for Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha, China. [4]Department of Materials Science and Engineering, The University of Texas at Austin, Austin, TX, USA. [5]California NanoSystems Institute, University of California, Los Angeles, CA, USA. *e-mail: yhuang@seas.ucla.edu; xduan@chem.ucla.edu

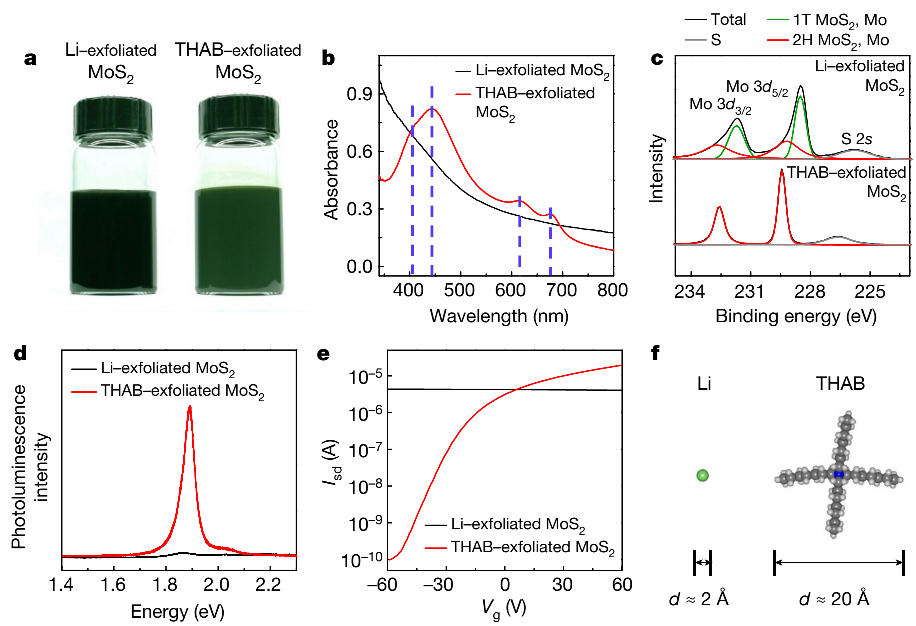


Fig. 2 | Exfoliation of semiconducting 2H-phase MoS_2 nanosheets (THAB–exfoliated MoS_2). **a**, Photograph of THAB–exfoliated MoS_2 nanosheets in isopropanol and Li–exfoliated MoS_2 nanosheets in water. The Li–exfoliated MoS_2 nanosheets were prepared via lithium intercalation, followed by sonication and exfoliation in water. **b**, Ultraviolet–visible absorption spectra of THAB–exfoliated MoS_2 (red) and Li–exfoliated MoS_2 (black). The vertical dashed lines indicate the four absorption peaks in the spectrum of THAB–exfoliated MoS_2. **c**, XPS spectra of THAB–exfoliated MoS_2 and Li–exfoliated MoS_2, showing the pure 2H phase for THAB–exfoliated MoS_2 (bottom) and mostly the 1T phase for Li–exfoliated MoS_2 (top). The green, red, grey and black curves indicate Mo in 1T-phase MoS_2, Mo in 2H-phase MoS_2, S and the overall signal, respectively. **d**, Photoluminescence spectra from a THAB–exfoliated MoS_2 monolayer (red) and a Li–exfoliated MoS_2 monolayer (black), after TFSI treatment. **e**, I_{sd}–V_g transfer characteristics of field-effect transistors made from a single THAB–exfoliated MoS_2 nanosheet (red) and from a single Li–exfoliated MoS_2 nanosheet (black), on the 300-nm-thick SiO_2/Si substrate and with $V_{sd} = 1$ V. The semiconducting 2H-phase THAB–exfoliated MoS_2 exhibits a much higher on/off ratio than the metallic 1T-phase Li–exfoliated MoS_2. **f**, The structures of THAB and Li, highlighting the size difference between the two, which results in distinct levels of electron injection into the MoS_2 crystal during the intercalation process.

negative electrochemical potential, THA^+ cations were inserted into the MoS_2 layer, causing a substantial volume expansion of the MoS_2 crystal (Extended Data Fig. 1c), similar to $N_2H_5^+$-intercalated MoS_2[21]. X-ray diffraction (XRD) demonstrates an increase in interlayer spacing from the original 6.1 Å to 22.9 Å (Extended Data Fig. 2a).

The expanded MoS_2 crystal was immediately sonicated in the polyvinylpyrrolidone solution in dimethylformamide (PVP/DMF) (Extended Data Fig. 1b), which rapidly exfoliated the intercalated compound into thin MoS_2 nanosheets and produced a greenish dispersion within several minutes. The PVP serves as a stabilizing agent to minimize restacking of the MoS_2 nanosheets. The resultant nanosheets were washed repeatedly to remove the excess PVP and large aggregates, and then dispersed in isopropanol to formulate a stable and easy-to-handle MoS_2 ink solution (Fig. 1a). The concentration of the MoS_2 nanosheets (up to around 10 mg ml^{-1}) could be tuned by centrifugation and re-dispersion in proper solvents for specific applications. The concentrated solution appears green and the diluted MoS_2 dispersion appears green and yellow (Fig. 1a). The colour of the dispersion is indicative of the formation of relatively thin, semiconducting nanosheets, because bulk or metallic MoS_2 crystals usually have strong featureless absorption in the entire visible-light range and thus appear black[22,23].

Atomic force microscopy (AFM) reveals that the exfoliated MoS_2 nanosheets have a thickness distribution of 3.8 ± 0.9 nm (mean ± standard deviation) and lateral dimensions of 0.5–2 μm (Fig. 1b–d). The Raman signal of the E^1_{2g} and A_{1g} peaks at 383.6 cm^{-1} and 408.3 cm^{-1} (Fig. 1e) indicates the few-layer (more than three) structure of the MoS_2 nanosheets[24], consistent with the AFM results. The intensity ratio of A_{1g}/E_{2g} is roughly 0.9, indicative of the absence of a large number of defects and of the high quality of the nanosheets[25]. Energy-dispersive X-ray spectroscopy (EDS) confirms the expected Mo and S ratio (Fig. 1f). The thin, flake-like nature of the MoS_2 nanosheets is also evident by the low contrast in the transmission electron microscopy (TEM) image (Fig. 1g). The hexagonal diffraction spots (Fig. 1h) and the lattice-resolved TEM image (Fig. 1i) indicate the high crystallinity of the MoS_2 nanosheets.

The THAB-intercalated, exfoliated ('THAB–exfoliated') MoS_2 nanosheets retain the intrinsic semiconducting 2H crystal phase, which differs fundamentally from the conventional Li-intercalated, exfoliated ('Li–exfoliated') MoS_2 nanosheets. The greenish colour of the dispersion of THAB–exfoliated MoS_2 nanosheets (Fig. 2a) signifies partial absorption in the visible range, as confirmed by the ultraviolet–visible absorption spectrum (Fig. 2b), and is an indication of the semiconducting behaviour of the nanosheets[11]. By contrast, the dispersion of Li–exfoliated MoS_2 nanosheets appears black, which indicates complete and featureless absorption in the visible range and the metallic nature of the nanosheets. The pure 2H phase of THAB–exfoliated MoS_2 nanosheets is further supported by X-ray photoelectron spectroscopy (XPS; Fig. 2c, Extended Data Fig. 3a–c); by contrast, the 1T phase is dominant in the Li–exfoliated MoS_2 nanosheets[5]. Furthermore, the THAB–exfoliated MoS_2 monolayer exhibits prominent photoluminescence that is orders of magnitude stronger than that of the Li–exfoliated MoS_2 monolayer (Fig. 2d). Owing to the retained semiconducting 2H phase, a field-effect transistor consisting of a single THAB–exfoliated nanosheet exhibited an electron mobility of about 10 $cm^2\ V^{-1}\ s^{-1}$ and current modulation over five orders of magnitude (Fig. 2e, Extended Data Fig. 3d, e). By contrast, the Li–exfoliated MoS_2 nanosheet displays negligible current modulation and a low field-effect mobility of 0.1 $cm^2\ V^{-1}\ s^{-1}$, consistent with the previous results[2,12].

As mentioned above, the larger size of THAB molecules compared to Li^+ ions (Fig. 2f) reduces the number of ions that fit into the gaps between MoS_2 layers and thus the number of electrons that are injected into the MoS_2 crystal. The atomic ratio of intercalated THA^+ is relatively small (roughly 2%; Extended Data Fig. 3f), corresponding to 0.02 electrons per MoS_2 formula unit, which is well below the

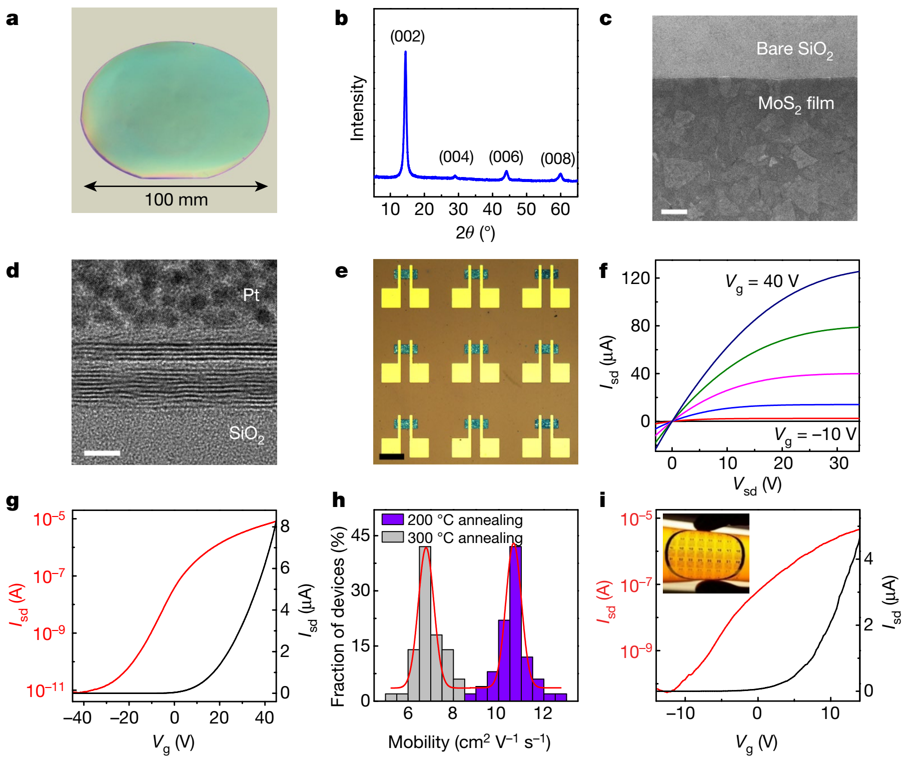


Fig. 3 | Large-scale solution-processable thin-film transistors. **a**, Photograph of the MoS_2 thin film deposited on a standard 100-mm-diameter SiO_2/Si wafer. **b**, XRD pattern of the MoS_2 thin film, demonstrating well-controlled assembly along the (001) direction. **c**, High-magnification SEM image of the MoS_2 thin film, showing the conformal stacking of individual nanosheets along the (001) direction. Scale bar, 1 μm. Only the top layer of nanosheets is visible; the lower layers appear as the dark background. **d**, Cross-sectional TEM image of the film deposited on the SiO_2/Si substrate, showing the broad-area plane-to-plane contacts between MoS_2 nanosheets. The top platinum (Pt) layer is deposited for the focused-ion beam process. The interlayer spacing (6.1 Å) obtained from both XRD and TEM studies agrees well with the pristine MoS_2 crystal structure and thus confirms the complete removal of intercalant. Scale bar, 5 nm. **e**, Optical microscope image of an array of back-gate thin-film transistors fabricated on the 90-nm-thick SiO_2/Si substrate. Scale bar, 100 μm. **f**, I_{sd}–V_{sd} output characteristics of the device fabricated on the 90-nm-thick SiO_2/Si substrate, under various gate biases from $V_g = 40$ V (blue) to $V_g = -10$ V (black). **g**, I_{sd}–V_g transfer characteristics of the same device as in **f**, with $V_{sd} = 1$ V. **h**, The statistical distribution of mobility for 50 individual transistors annealed at 200 °C (purple bars) and at 300 °C (grey bars). The red curves are Gaussian fits. **i**, MoS_2 thin-film transistors fabricated on a flexible substrate, demonstrating similar behaviour to that of the devices fabricated on the rigid SiO_2/Si substrate. $V_{sd} = 10$ V. Inset, photograph of flexible devices on plastic.

phase-transition threshold (0.29 electrons per MoS_2 formula unit). Together, our systematic characterizations and analyses demonstrate unambiguously that we have produced a dispersion of phase-pure semiconducting 2H-MoS_2 nanosheets, which serves as a high-quality semiconductor ink that is indispensable for solution-processable large-area MoS_2 thin-film electronics.

With the formulation of the stable ink, large-area thin films could be prepared on diverse substrates using various solution-processing approaches. For example, we deposited visually uniform MoS_2 thin films on a 100-mm-diameter SiO_2/Si wafer using a spin-coating process (Fig. 3a). Alternatively, the ink solution may be applied using the industrial roll-to-roll coating process to produce thin films of even larger area. After acid cleaning and moderate thermal annealing (200–300 °C) to remove residual PVP ligands (see Methods), we obtained high-quality 2D semiconducting thin films, which can be further processed using standard photolithography to create desired patterns for the fabrication of arrays of devices (Extended Data Fig. 4a). The final MoS_2 thin films are free from organic contamination by THAB (intercalant) or PVP (surfactant), as evidenced by the lack of any N signal in EDS (Fig. 1f) or XPS (Extended Data Fig. 3c) spectra.

XRD analysis of the resulting thin film displays only {001} planes (Fig. 3b), indicative of the well-controlled assembly of MoS_2 nanosheets along the *c* axis of the crystal. High-resolution scanning electron microscopy (SEM) further confirms that the nanosheets lie flat on the surface of the substrate (Fig. 3c). The thin film consists of roughly two to three layers of stacked, few-layer nanosheets with conformal plane-to-plane contacts and an average thickness of around 10 nm (Extended Data Fig. 4b, c). The ultrathin and uniform nature of the THAB–exfoliated MoS_2 nanosheets is essential for the formation of compact thin films, in contrast to thin films prepared from liquid-exfoliated materials, which consist of randomly packed nanosheets owing to the much larger thickness distribution of the nanosheets[14].

Cross-sectional TEM demonstrates that the broad-area plane-to-plane contacts between the dangling-bond-free nanosheets have many regions that are nearly indistinguishable from the natural van der Waals interfaces between the atomic layers of MoS_2 (Fig. 3d, Extended Data Fig. 4d, e). Such high-quality dangling-bond-free contacts with pinning-free van der Waals interfaces ensure optimized charge transport between individual nanosheets within the thin film, which is essential for achieving excellent electrical performance in thin-film electronics[13]. By contrast, conventional nanostructured thin films consisting of semiconductor quantum dots or nanowires are typically plagued by dangling bonds and chemical disorder at grain boundaries[15].

To explore the potential of the thin films for electronic applications, we fabricated back-gate MoS_2 thin-film transistors on the 90-nm-thick SiO_2/Si substrate using standard photolithography and etching processes (Fig. 3e, Extended Data Fig. 4). The I_{sd}–V_{sd} output curves (where I_{sd} is the source–drain current and V_{sd} is the source–drain bias; Fig. 3f) and I_{sd}–V_g transfer curves (where V_g is the gate voltage; Fig. 3g) are characteristic of a typical n-type transistor. On the basis of the transfer characteristics, we derived an average mobility of around 7–11 $cm^2 V^{-1} s^{-1}$ and an on/off ratio of 10^6 (Fig. 3h, Extended Data Fig. 5, 6). The mobility achieved in the film annealed at 300 °C is slightly

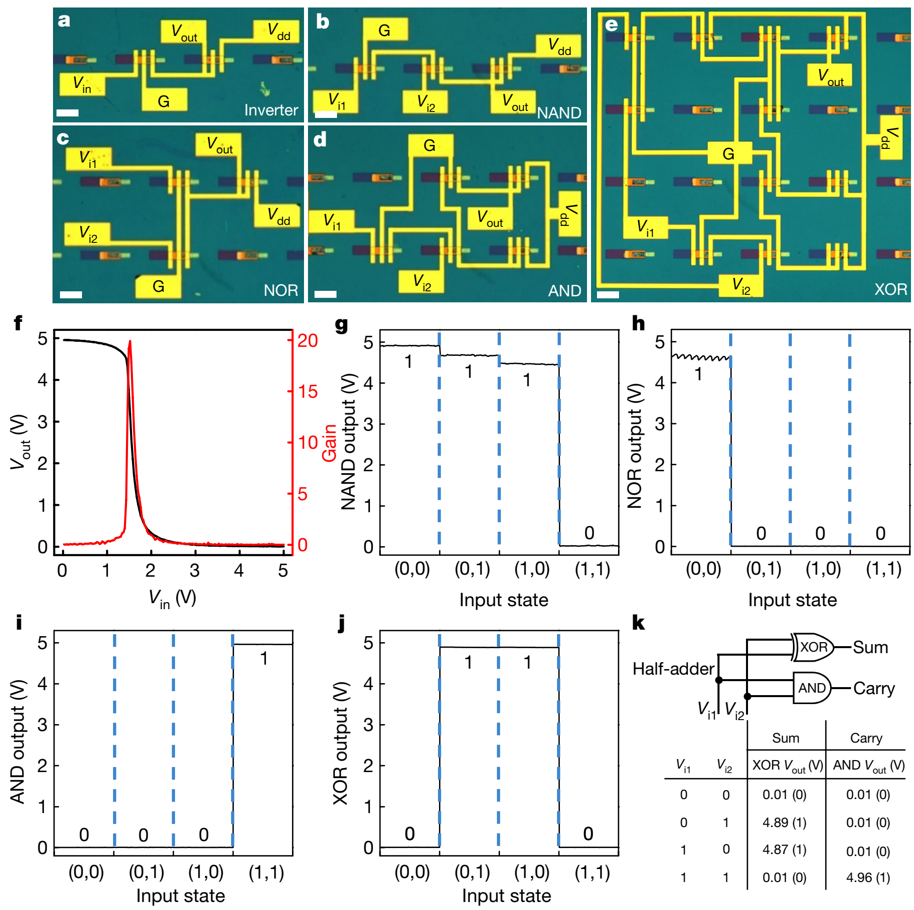


Fig. 4 | Logic gates and computational circuits from solution-processable MoS_2 thin-film transistors. a–e, Optical images of an inverter (**a**) and NAND (**b**), NOR (**c**), AND (**d**) and XOR (**e**) gates. Scale bars, 100 μm. G, ground; V_{dd}, drain supply voltage; V_{in}, input voltage; V_{out}, output voltage; V_{i1}, input voltage 1; V_{i2}, input voltage 2. **f**, Output voltage V_{out} (black, left axis) and voltage gain (red, right axis) of the integrated MoS_2 logic inverter as a function of the input voltage V_{in}, highlighting a high voltage gain of roughly 20. **g–j**, Output voltage of the logic NAND (**g**), NOR (**h**), AND (**i**) and XOR (**j**) gates at four typical input states (V_{i1}, V_{i2}) (separated by blue dashed vertical lines), with a power supply of $V_{dd} = 5$ V. '0' and '1' labelled in the plot represent the low and high binary output states, respectively. **k**, Experimental truth table for the logic half-adder. The logic half-adder was obtained by using an XOR gate as the 'sum' and an AND gate as the 'carry'. '0' or '1' in parentheses indicates the corresponding binary output state.

lower than that achieved in the film annealed at 200 °C, which may be attributed to the fact that the transistor is not fully turned on to reach the maximum transconductance and the peak mobility within the applied gate voltage range (Extended Data Fig. 7). The higher annealing temperature reduces unintentional impurity doping, lowers the electron concentration and shifts the threshold voltage towards more positive values, making the achieved mobility further away from the peak value.

The mobility values achieved in solution-processed thin films are similar to that achieved in the single MoS_2 nanosheet (about 10 $cm^2\ V^{-1}\ s^{-1}$). This finding indicates that the contacts between MoS_2 nanosheets in the large-area thin film do not greatly hinder the transport properties of the thin film, owing to the optimized charge transport across the stacked nanosheets due to the broad-area dangling-bond-free plane-to-plane contacts between them. Such contacts are advantageous compared with the point-to-point contacts in zero-dimensional quantum-dot thin films and in one-dimensional nanowire thin films, which, as mentioned previously, have considerable interfacial dangling bonds and chemical disorder[13,15]. The carrier mobility achieved in our large-area 2D thin films is much higher than the carrier mobilities reported in previous solution-processed MoS_2 thin films (less than 0.4 $cm^2\ V^{-1}\ s^{-1}$)[12,14,16] and of the same order of magnitude as those reported in wafer-scale polycrystalline thin films obtained by chemical vapour deposition (CVD; 14–29 $cm^2\ V^{-1}\ s^{-1}$)[7,26,27]. A comprehensive comparison is provided in Extended Data Table 1. With a stable ink solution and a relatively low processing temperature, thin-film devices can also readily be prepared on flexible plastic substrates with comparable electronic performance (Fig. 3i), which demonstrates the potential of the 2D semiconductor ink for flexible and wearable electronics[28–30].

The ability to process uniform, large-area semiconducting thin films from the phase-pure 2H-MoS_2 ink enables the scalable fabrication of thin-film transistors with a high yield (greater than 95%) and allows us to create more complex devices, such as logic gates and computational circuits. To this end, we constructed MoS_2 thin-film transistors with a local back gate that is insulated from the MoS_2 thin film by an underlying 30-nm-thick Al_2O_3 dielectric layer (Extended Data Fig. 8a–c). Next, we constructed a logic inverter using two locally gated MoS_2 thin-film transistors (Fig. 4a, Extended Data Fig. 8d). The voltage swing of the resulting device is indicative of an inverter function, with a high output at low input and a low output at high input, demonstrating a substantial voltage gain of about 20 (Fig. 4f). Such a high gain is crucial for signal propagation and logic operations in integrated circuits. We also designed and fabricated logic NAND, NOR, AND and XOR gates by integrating 3, 3, 5 and 11 thin-film transistors, respectively (Fig. 4b–e, Extended Data Fig. 8e–h), and achieved the desired logic function (Fig. 4g–j). The successful realization of these diverse logic functions allows us to further construct computational circuits, such as a half-adder (Fig. 4k), which corresponds to the addition of two one-bit binary numbers. The logic half-adder uses the logic XOR gate as the 'sum', with an additional logic AND gate as the 'carry'. The data summarized in the experimental truth table (Fig. 4k) clearly demonstrate that we have successfully implemented a basic logic computation.

The creation of these basic logic gates enables the organization of virtually any digital integrated circuit and opens up a scalable pathway to high-performance logic applications using solution-processable 2D semiconductor inks. Because the integration of complex logic circuits necessitates the scalable and high-yield fabrication of multiple high-performance thin-film transistors (for example, with on/off ratios of more than 10^3 and inverter gains of more than 1), it has been challenging to achieve with thin films of nanosheets prepared using lithium intercalation and exfoliation or liquid exfoliation (which have limited carrier mobility, low on/off ratio or poor film quality). By contrast, we have successfully created integrated computational circuits from solution-processable 2D semiconductor thin films (beyond a single-transistor or few-transistor NAND or NOR gate on individual

mechanically exfoliated or CVD-grown MoS_2 nanosheets). Our results clearly highlight the quality, uniformity and scalability of the solution-processed MoS_2 thin films. Furthermore, our quaternary ammonium intercalation and exfoliation strategy can be applied generally to a wide range of 2D materials, including WSe_2, Bi_2Se_3, $NbSe_2$, In_2Se_3, Sb_2Te_3 and black phosphorus (Extended Data Fig. 9), to establish a library of 2D-material inks with diverse functionalities. Our study thus provides a robust pathway to the scalable production of high-quality nanosheets for large-area electronics, optoelectronics and thermoelectrics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0574-4.

Received: 7 February 2018; Accepted: 13 August 2018;
Published online 3 October 2018.

1. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS_2 transistors. *Nat. Nanotechnol*. **6**, 147–150 (2011).
2. Zeng, Z. et al. Single-layer semiconducting nanosheets: high-yield preparation and device fabrication. *Angew. Chem. Int. Ed*. **50**, 11093–11097 (2011).
3. Wang, Q. H., Kalantar-Zadeh, K., Kis, A., Coleman, J. N. & Strano, M. S. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. *Nat. Nanotechnol*. **7**, 699–712 (2012).
4. Butler, S. Z. et al. Progress, challenges, and opportunities in two-dimensional materials beyond graphene. *ACS Nano* **7**, 2898–2926 (2013).
5. Kappera, R. et al. Phase-engineered low-resistance contacts for ultrathin MoS_2 transistors. *Nat. Mater*. **13**, 1128–1134 (2014).
6. Cheng, R. et al. Few-layer molybdenum disulfide transistors and circuits for high-speed flexible electronics. *Nat. Commun*. **5**, 5143 (2014).
7. Kang, K. et al. High-mobility three-atom-thick semiconducting films with wafer-scale homogeneity. *Nature* **520**, 656–660 (2015).
8. Gao, M.-R. et al. An efficient molybdenum disulfide/cobalt diselenide hybrid catalyst for electrochemical hydrogen generation. *Nat. Commun*. **6**, 5982 (2015).
9. Desai, S. B. et al. MoS_2 transistors with 1-nanometer gate lengths. *Science* **354**, 99–102 (2016).
10. Lin, Z. et al. Scalable solution-phase epitaxial growth of symmetry-mismatched heterostructures on two-dimensional crystal soft template. *Sci. Adv*. **2**, e1600993 (2016).
11. Eda, G. et al. Photoluminescence from chemically exfoliated MoS_2. *Nano Lett*. **11**, 5111–5116 (2011).
12. He, Q. et al. Fabrication of flexible MoS_2 thin-film transistor arrays for practical gas-sensing applications. *Small* **8**, 2994–2999 (2012).
13. Lin, Z. et al. Solution processable colloidal nanoplates as building blocks for high-performance electronic thin films on flexible substrates. *Nano Lett*. **14**, 6547–6553 (2014).
14. Kelly, A. G. et al. All-printed thin-film transistors from networks of liquid-exfoliated nanosheets. *Science* **356**, 69–73 (2017).
15. Talapin, D. V., Lee, J.-S., Kovalenko, M. V. & Shevchenko, E. V. Prospects of colloidal nanocrystals for electronic and optoelectronic applications. *Chem. Rev*. **110**, 389–458 (2010).
16. Xi, Y. et al. Fabrication of MoS_2 thin film transistors via selective-area solution deposition methods. *J. Mater. Chem. C* **3**, 3842–3847 (2015).
17. Coleman, J. N. et al. Two-dimensional nanosheets produced by liquid exfoliation of layered materials. *Science* **331**, 568–571 (2011).
18. Wang, Y. et al. Structural phase transition in monolayer $MoTe_2$ driven by electrostatic doping. *Nature* **550**, 487–491 (2017).
19. Li, Y., Duerloo, K.-A. N., Wauson, K. & Reed, E. J. Structural semiconductor-to-semimetal phase transition in two-dimensional materials induced by electrostatic gating. *Nat. Commun*. **7**, 10671 (2016).
20. Wang, C. et al. Monolayer atomic crystal molecular superlattices. *Nature* **555**, 231–236 (2018).
21. Zheng, J. et al. High yield exfoliation of two-dimensional chalcogenides using sodium naphthalenide. *Nat. Commun*. **5**, 2995 (2014).
22. O'Neill, A., Khan, U. & Coleman, J. N. Preparation of high concentration dispersions of exfoliated MoS_2 with increased flake size. *Chem. Mater*. **24**, 2414–2421 (2012).
23. Halim, U. et al. A rational design of cosolvent exfoliation of layered materials by directly probing liquid-solid interaction. *Nat. Commun*. **4**, 2213 (2013).
24. Lee, C. et al. Anomalous lattice vibrations of single-and few-layer MoS_2. *ACS Nano* **4**, 2695–2700 (2010).
25. Parkin, W. M. et al. Raman shifts in electron-irradiated monolayer MoS_2. *ACS Nano* **10**, 4134–4142 (2016).
26. Zhao, M. et al. Large-scale chemical assembly of atomically thin transistors and circuits. *Nat. Nanotechnol*. **11**, 954–959 (2016).
27. Lim, Y. R. et al. Wafer-scale, homogeneous MoS_2 layers on plastic substrates for flexible visible-light photodetectors. *Adv. Mater*. **28**, 5025–5030 (2016).
28. Lin, Z. et al. Cosolvent approach for solution-processable electronic thin films. *ACS Nano* **9**, 4398–4405 (2015).
29. Sun, Y. & Rogers, J. A. Inorganic semiconductors for flexible electronics. *Adv. Mater*. **19**, 1897–1916 (2007).
30. Park, S., Vosguerichian, M. & Bao, Z. A review of fabrication and applications of carbon nanotube film-based flexible electronics. *Nanoscale* **5**, 1727–1752 (2013).

Acknowledgements Xiangfeng Duan acknowledges support from the US Department of Energy, Office of Basic Energy Sciences, Division of Materials Science and Engineering through award DE-SC0018828. (material synthesis and structural characterizations). Y.H. acknowledges financial support from National Science Foundation grant EFRI-1433541 (thin film device fabrication and characterizations). We acknowledge the Electron Imaging Center at UCLA for TEM technical support and the Nanoelectronics Research Facility at UCLA for device fabrication technical support.

Reviewer information *Nature* thanks K. Xiao and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Xiangfeng Duan designed the research. Z.L. developed the intercalation–exfoliation approach, fabricated the electronic devices and analysed the results. Z.L., U.H. and Yuanyue Liu solved the structure of the intercalated crystal. Yuan Liu, M.D., Y.W., C.J., P.C. and Xidong Duan assisted with the fabrication of the electronic device, the measurements and the analysis of the results. Z.L., Y.W. and Chen Wang collected the Raman and photoluminescence spectra. Z.L., F.S., M.L. and Chengzhang Wan performed the electron microscopy analyses on the exfoliated materials. Xiangfeng Duan and Y.H. supervised the research. Xiangfeng Duan and Z.L. co-wrote the paper. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0574-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0574-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.H. and X.D.

METHODS

Intercalation of 2D layered crystals with quaternary ammonium bromide. A two-electrode electrochemical cell was used to conduct the intercalation reaction. A thin piece of cleaved MoS_2 crystal and graphite rod were placed as the cathode and anode, respectively. The MoS_2 piece was anchored onto the copper wire with a conductive copper tape or fixed directly by an alligator clip. Quaternary ammonium bromide (tetraethylammonium bromide, tetrapropylammonium bromide, tetrabutylammonium bromide, tetraheptylammonium bromide or tetradecylammonium bromide; 98% from TCI) solution in acetonitrile (40 ml; 5 mg ml^{-1} or higher) served as the electrolyte. The applied voltage was set to be 5–10 V and the intercalation process was allowed to proceed for 1 h. The electrolyte concentration and intercalation voltage could be tuned to adjust the reaction rate. During the intercalation process, a negative voltage was applied to the MoS_2 (cathode) to insert the positively charged THA^+ (tetraheptylammonium cation) into the crystal. The intercalation of THA^+ has the same mechanism as that of Li^+ intercalation, despite the larger size of THA^+. In analogy to Li^+ intercalation, the insertion of each THA^+ requires the injection of one electron into the MoS_2 host crystal from the external circuit to balance the charge:

$$MoS_2 + xTHA^+ + xe^- \rightarrow (THA^+)_x MoS_2^{x-} \quad (1)$$

At the same time, Br^- is oxidized to Br_2 on the graphite side (anode):

$$xBr^- \rightarrow (x/2)Br_2 + xe^- \quad (2)$$

as was confirmed by the formation of a yellow solution surrounding the graphite electrode with the progression of intercalation. After the reaction completed, the MoS_2 piece evolved into an expanded and fluffy structure (Extended Data Fig. 1c).

The interlayer distance expansion could be modulated by selecting molecules of various sizes. For example, the interlayer distance evolves from 10.9 Å to 14.3 Å, 18.5 Å and 24.2 Å in MoS_2 intercalated with tetraethylammonium bromide, tetrapropylammonium bromide, tetrabutylammonium bromide and tetradecylammonium bromide, respectively (Extended Data Fig. 2). Therefore, the dimension of molecules may be used to tailor the structure (such as interlayer spacing and molecule packing density) and properties (electron-doping level) of intercalated compounds. For the intercalation of other layered materials, a similar procedure was carried out, but with the corresponding crystals used as the cathode.

The Li–exfoliated MoS_2 nanosheets were prepared using butyllithium intercalation (1.6 M solution in hexane; Sigma-Aldrich) and sonication-assisted exfoliation in water[11]. The repeated wash and purification process was carried out to obtain the final Li–exfoliated MoS_2 nanosheets in water.

Formulation of ink solutions. The as-intercalated piece was rinsed with absolute ethanol before sonication in 40 ml 0.2 M PVP/DMF solution (PVP: molecular weight of about 40,000, Sigma-Aldrich) for 30 min. The fluffy bulk material was manually broken down into smaller pieces to facilitate the sonication-assisted exfoliation. For MoS_2, a greenish dispersion of nanosheets formed after sonication for several minutes, indicating the successful exfoliation of the intercalated compound. The dispersion was subsequently centrifuged and washed with isopropanol (IPA) twice more to remove excessive PVP. To get rid of large chunks or other impurities, the final dispersion in IPA was centrifuged at 1,000 r.p.m. for 3 min and precipitates were discarded. Ink solutions could be made in ethanol, DMF or other solvents for specific applications.

Thin-film deposition and post-treatment. The MoS_2–IPA ink solution was used for the film deposition on the SiO_2/Si, glass slide or plastic substrate. An additional centrifugation at 3,000 r.p.m. was performed for 3 min before adjusting the concentration of the final ink solution. The optical absorbance was used to monitor the concentration of the MoS_2 dispersion. For a standard ink solution, the peak absorbance at around 440 nm was tuned to be 0.70 (cuvette length of 4 mm) for the solution that was diluted by a factor of 50. Then, the ink solution was further concentrated from the original 1.6 ml to 0.6 ml for deposition. The ink solution was spin-coated three times on the 90-nm-thick SiO_2/Si substrate (three or four times for the polyimide substrate) at a speed of 2,000 r.p.m. for 20 s. The SiO_2/Si and plastic substrates were pre-cleaned with IPA and treated with oxygen plasma (about 5 min) before spin coating. If the film is too thick, then the fabricated transistors may not be completely turned off owing to the limited gate modulation depth. Therefore, the ink concentration and spin-coating procedure need to be optimized for different types of substrate to produce a continuous and reasonably thick film to achieve the desired performance. For deposition on a polyimide substrate, a layer of polyimide solution (Sigma-Aldrich) was first coated to cure the surface. Then, the thin film prepared on various substrates was treated with 10 mg ml^{-1} bis(trifluoromethane)sulfonimide (TFSI; more than 95.0%, Sigma-Aldrich) in 1,2-dichloroethane (Fisher Chemical) at 80 °C for 1 h[31], before the final thermal annealing in a tube furnace for 1 h (in argon or argon/hydrogen atmosphere; ramp rate of 12 °C min^{-1}).

Device fabrication. Thin-film transistors were fabricated on various substrates (such as 90-nm-thick SiO_2/Si) following standard photolithography, drying etching and e-beam evaporation of Ti/Au (30 nm/50 nm) source/drain electrodes. For the single-nanosheet device, e-beam lithography was used to define source and drain electrodes on the SiO_2/Si substrate with a 300-nm-thick oxide layer. In devices fabricated on the plastic substrate, a 60-nm top-gate Al_2O_3 dielectric layer was deposited using the standard atomic-layer deposition at a processing temperature of 100 °C. To fabricate the logic gates, Ti/Au (10 nm/40 nm) local gate electrodes were defined on the substrate, followed by ALD deposition of a 30-nm Al_2O_3 dielectric layer. MoS_2 ink solution was coated on the substrate to obtain films at the desired thickness. TFSI and thermal annealing treatments (400 °C) were performed before patterning the MoS_2, opening the gate electrode window and defining the top-contact electrodes for the completion of circuits.

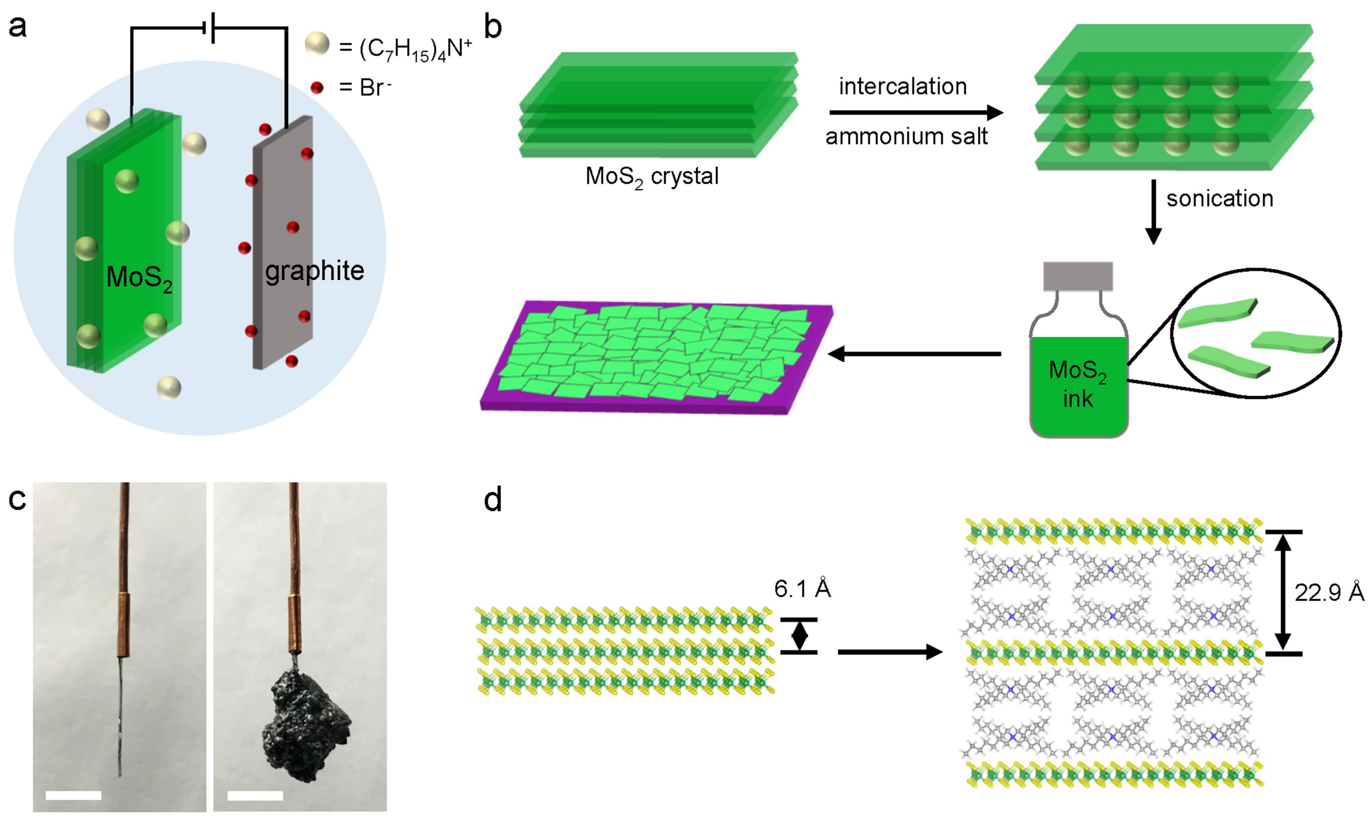
Characterizations. Characterizations were carried out using SEM (JEOL JSM-6700F FE-SEM) with EDS (EDAX), TEM (T12 Quick CryoEM and CryoET FEI: acceleration voltage, 120 KV; Titan S/TEM FEI: acceleration voltage, 300 KV), XRD (Panalytical X'Pert Pro X-ray Powder Diffractometer), AFM (Bruker Dimension Icon Scanning Probe Microscope), UV–vis–NIR spectroscopy (Shimadzu 3100 PC), Raman and photoluminescence spectroscopy (Horiba, 488-nm laser wavelength), and XPS (AXIS Ultra DLD). To obtain photoluminescence spectra, exfoliated MoS_2 monolayers after the TFSI treatment were used. The measurements of the transport characteristic were conducted at room temperature under ambient conditions (in vacuum and dark) with a probe station and a computer-controlled analogue-to-digital converter.

DFT calculations. Density functional theory (DFT) calculations were conducted using the Vienna Ab-initio Simulation Package (VASP), with projector augmented wave (PAW) pseudopotentials, the Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional, and 400 eV for the plane-wave cut-off. We used a 3 nm × 3 nm × 3 nm box to model $(C_7H_{15})_4N^+$, which was fully relaxed until the final force on each atom was less than 0.01 eV $Å^{-1}$. We tested several configurations and found that the most stable one has a planar structure as shown in Extended Data Fig. 1d, with a 'thickness' of about 6.1 Å. This structure minimizes the steric repulsion and resembles that of a tetra-*n*-butylammonium cation. Because the interlayer distance is expanded by 16.8 Å, we assume that two layers of $(C_7H_{15})_4N^+$ were intercalated, given an interlayer distance of roughly 4.6 Å (considering the van der Waals interaction and Coulomb repulsion between organic cations).

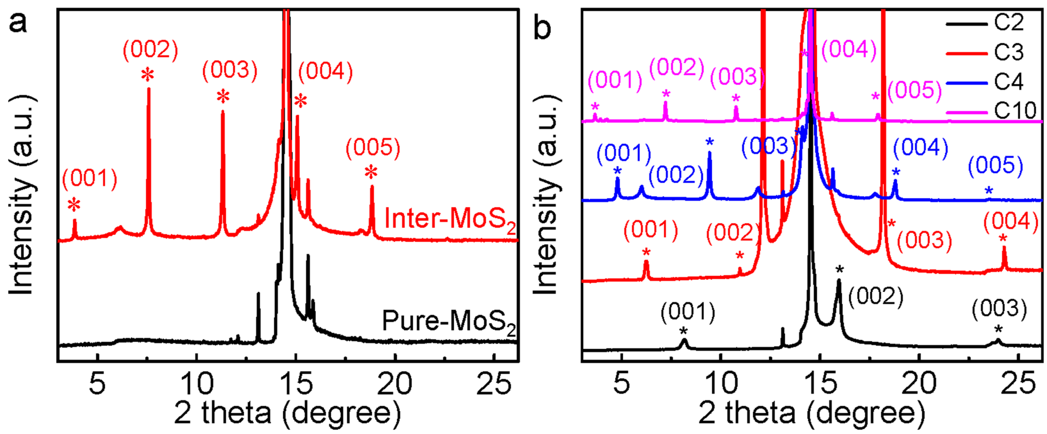
Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

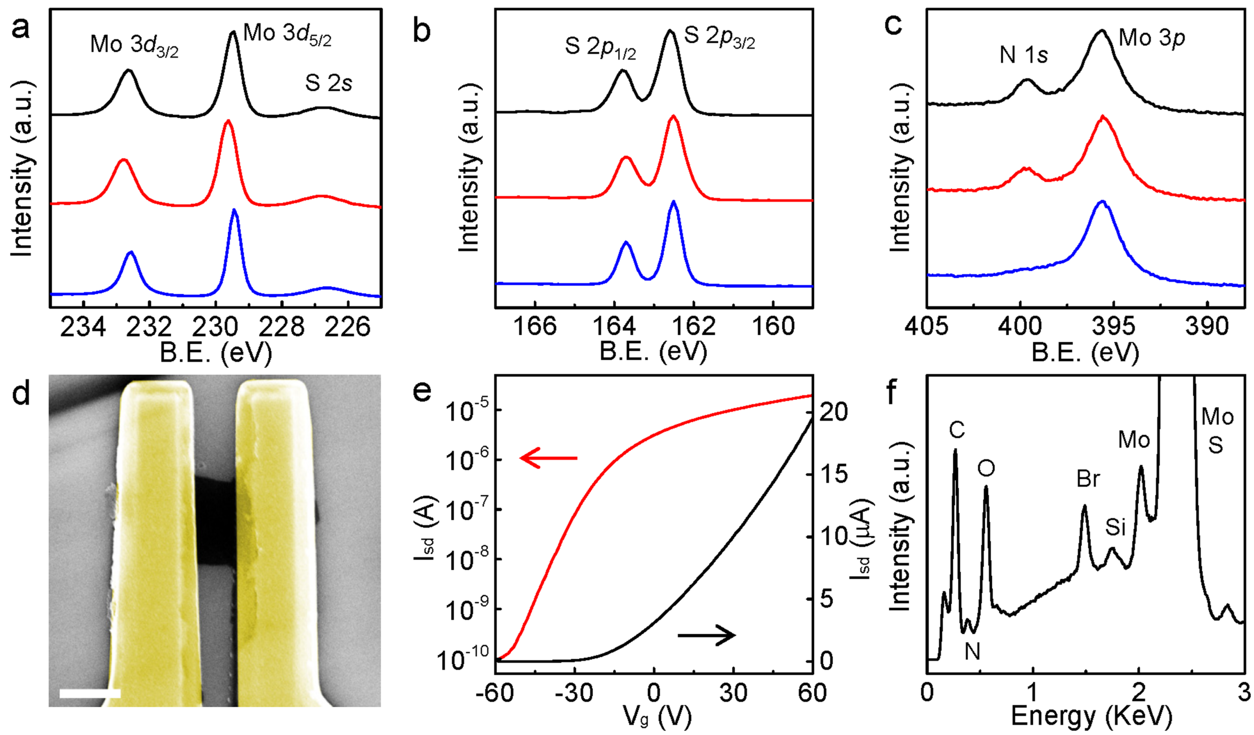
31. Amani, M. et al. Near-unity photoluminescence quantum yield in MoS_2. *Science* **350**, 1065–1068 (2015).
32. Yu, X., Prévot, M. S. & Sivula, K. Multiflake thin film electronic devices of solution processed 2D MoS_2 enabled by sonopolymer assisted exfoliation and surface modification. *Chem. Mater*. **26**, 5892–5899 (2014).
33. Li, J., Naiini, M. M., Vaziri, S., Lemme, M. C. & Östling, M. Inkjet printing of MoS_2. *Adv. Funct. Mater*. **24**, 6524–6531 (2014).
34. Kim, T.-Y. et al. Electrical properties of synthesized large-area MoS_2 field-effect transistors fabricated with inkjet-printed contacts. *ACS Nano* **10**, 2819–2826 (2016).
35. Tao, L. et al. Centimeter-scale CVD growth of highly crystalline single-layer MoS_2 film with spatial homogeneity and the visualization of grain boundaries. *ACS Appl. Mater. Interfaces* **9**, 12073–12081 (2017).
36. Gurarslan, A. et al. Surface-energy-assisted perfect transfer of centimeter-scale monolayer and few-layer MoS_2 films onto arbitrary substrates. *ACS Nano* **8**, 11522–11528 (2014).
37. Jeon, J. et al. Layer-controlled CVD growth of large-area two-dimensional MoS_2 films. *Nanoscale* **7**, 1688–1695 (2015).
38. Song, X. et al. A novel synthesis method for large-area MoS_2 film with improved electrical contact. *2D Mater*. **4**, 025051 (2017).
39. Lee, Y. et al. Trap-induced photoresponse of solution-synthesized MoS_2. *Nanoscale* **8**, 9193–9200 (2016).
40. George, A. S. et al. Wafer scale synthesis and high resolution structural characterization of atomically thin MoS_2 layers. *Adv. Funct. Mater*. **24**, 7461–7466 (2014).
41. Liu, K.-K. et al. Growth of large-area and highly crystalline MoS_2 thin layers on insulating substrates. *Nano Lett*. **12**, 1538–1544 (2012).
42. Schmidt, H. et al. Transport properties of monolayer MoS_2 grown by chemical vapor deposition. *Nano Lett*. **14**, 1909–1913 (2014).
43. Zheng, J. et al. High-mobility multilayered MoS_2 flakes with low contact resistance grown by chemical vapor deposition. *Adv. Mater*. **29**, 1604540 (2017).



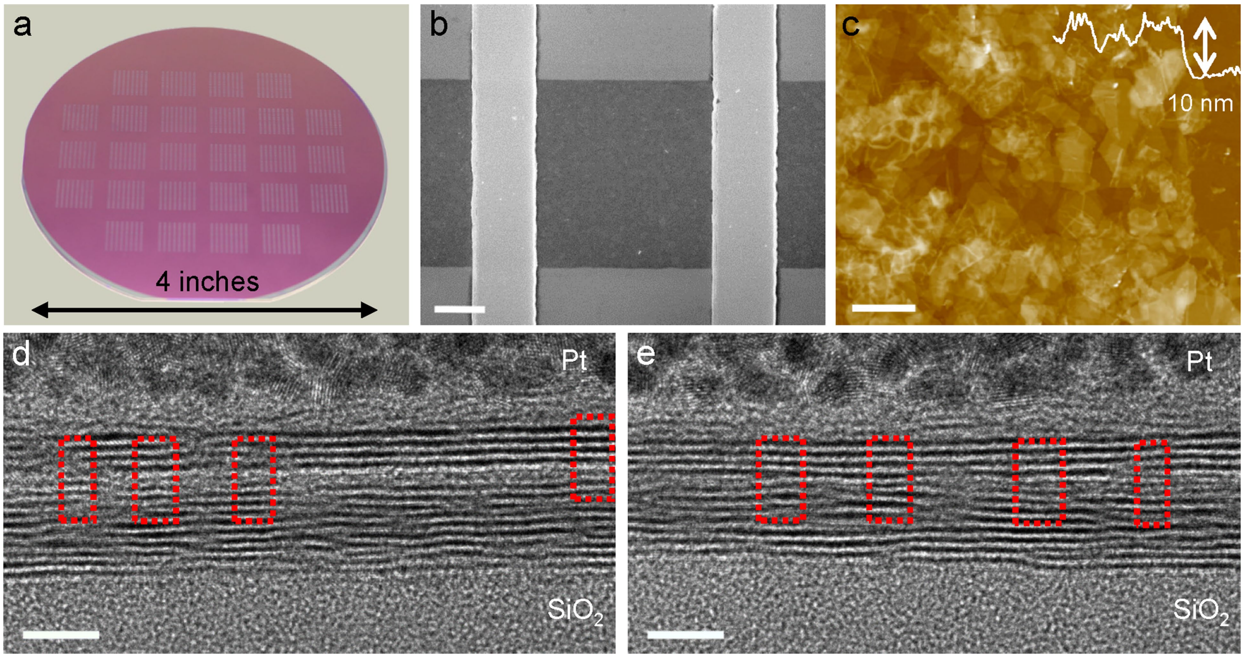
Extended Data Fig. 1 | Molecular intercalation and exfoliation of MoS_2 nanosheets. a, Schematic of the electrochemical intercalation of MoS_2 with THAB. **b**, Schematic of the intercalation and exfoliation of the MoS_2 crystal to produce 2D-nanosheet ink, which can be processed further into continuous large-area thin films with broad-area plane-to-plane contacts between stacked nanosheets. **c**, Photographs of a thin piece of MoS_2 crystal before (left) and after (right) THAB intercalation. The MoS_2 crystal expands by a factor of about 20 along the (001) direction. Scale bars, 5 mm. **d**, Schematic of the interlayer spacing expansion from the original 6.1 Å to 22.9 Å, caused by the intercalation of THAB molecules.



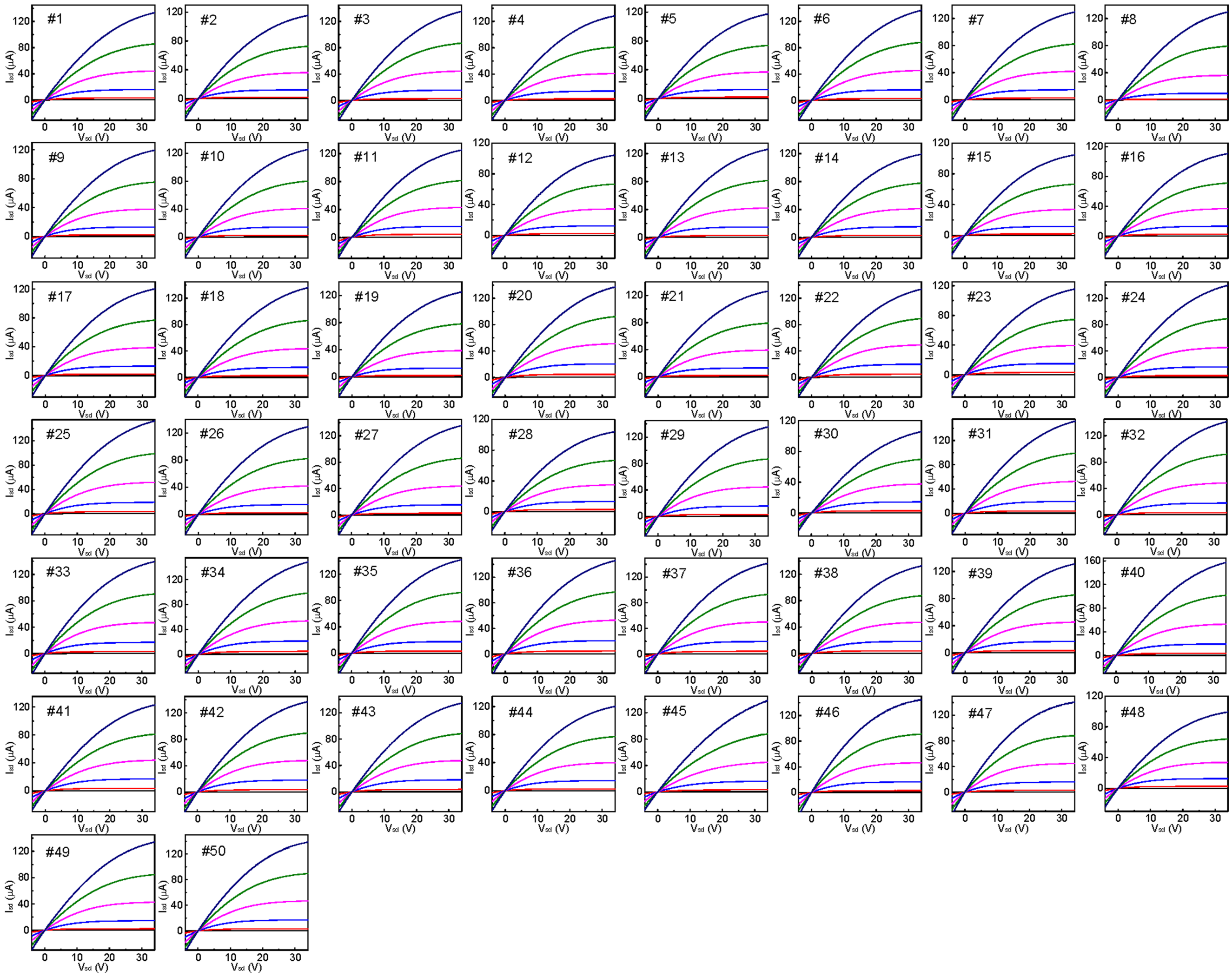
Extended Data Fig. 2 | XRD patterns of MoS_2 crystals intercalated with various ammonium molecules. **a**, XRD patterns of the THAB-intercalated (inter-MoS_2) and pure MoS_2. The MoS_2 interlayer spacing expands from the original 6.1 Å to 22.9 Å after the electrochemical intercalation of THAB. The peaks indicated with asterisks are for the intercalated MoS_2; the rest are from MoS_2 itself. For XRD, the MoS_2 crystal was allowed to be intercalated for a short period (about 1 min) to capture the intermediate state where the intercalation occurs but the expansion is not sufficient to cause severe shape deformation. **b**, XRD patterns for MoS_2 crystals intercalated with various ammonium molecules. C2, C3, C4, and C10 represent tetraethylammonium bromide ($(C_2H_5)_4NBr$; TEAB), tetrapropylammonium bromide ($(C_3H_7)_4NBr$; TPAB), tetrabutylammonium bromide ($(C_4H_9)_4NBr$; TBAB) and tetradecylammonium bromide ($(C_{10}H_{21})_4NBr$; TDAB), respectively. With the increasing molecule size (from C2 to C3, C4 and C10), the interlayer distance expansion is more substantial, ranging from 10.9 Å to 14.3 Å, 18.5 Å and 24.2 Å.



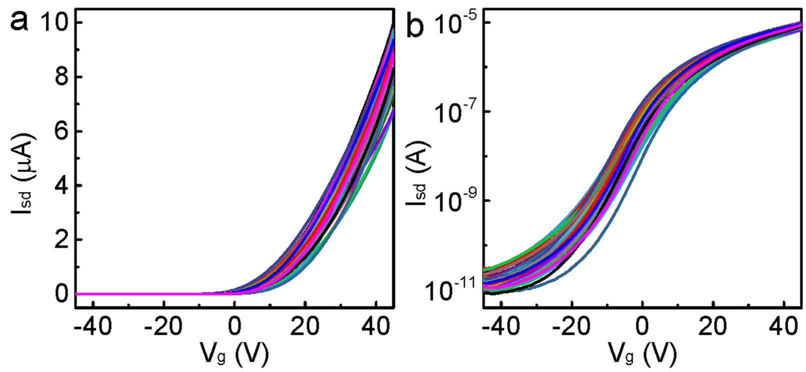
Extended Data Fig. 3 | Elemental and electrical analyses on exfoliated MoS_2 nanosheets. a–c, XPS spectra for the as-deposited (black), TFSI-treated (red), and TFSI-treated and annealed (blue) MoS_2 film, including the Mo 3*d* scan (**a**), S 2*p* scan (**b**) and N 1*s* scan (**c**). All three MoS_2 films have the same 2H crystal structure. The absence of any N signal after TFSI treatment and thermal annealing indicates the complete removal of ammonium molecules and PVP. a.u., arbitrary units. **d**, SEM image of a single-nanosheet device fabricated on a 300-nm-thick SiO_2/Si substrate. Electrodes are false-coloured. Scale bar, 1 μm. **e**, I_{sd}–V_g transfer characteristics of the device with $V_{sd} = 1$ V. **f**, EDS spectrum of the intercalated THAB–exfoliated MoS_2 superlattice. It shows the co-existence of Mo and S from MoS_2 and that of N and Br from THAB. The atomic ratio N/Br is approximately 2% of Mo/S. During the reaction, Br_2 was produced at the anode, as indicated by the emergence of the dark yellow Br_2 solution. However, after the electrochemical potential was withdrawn, the highly active Br_2 quickly back-reacted with the THA^+–MoS_2^- layers to form the final MoS_2–THAB superlattice structure. Consequently, N and Br elements were both detected in the THAB–exfoliated MoS_2 structure.



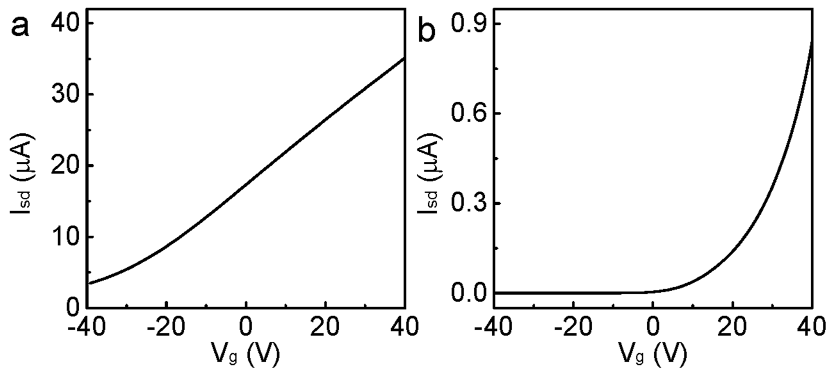
Extended Data Fig. 4 | Characterizations of MoS_2 films fabricated on a substrate. **a**, Photograph of the photolithographically patterned MoS_2 thin film on a standard 100-mm-diameter SiO_2/Si wafer. **b**, **c**, SEM image of the transistor fabricated with photolithography (**b**) and AFM analysis (**c**). Scale bars, 10 μm (**b**) and 2 μm (**c**). **d**, **e**, Cross-sectional TEM images of plane-to-plane contacts between MoS_2 nanosheets; **d** and **e** show two different regions of the thin film. The red dashed boxes indicate the regions where two nanosheets exhibit a contact that is nearly indistinguishable from a van der Waals interface between atomic layers of MoS_2. Scale bars, 5 nm.



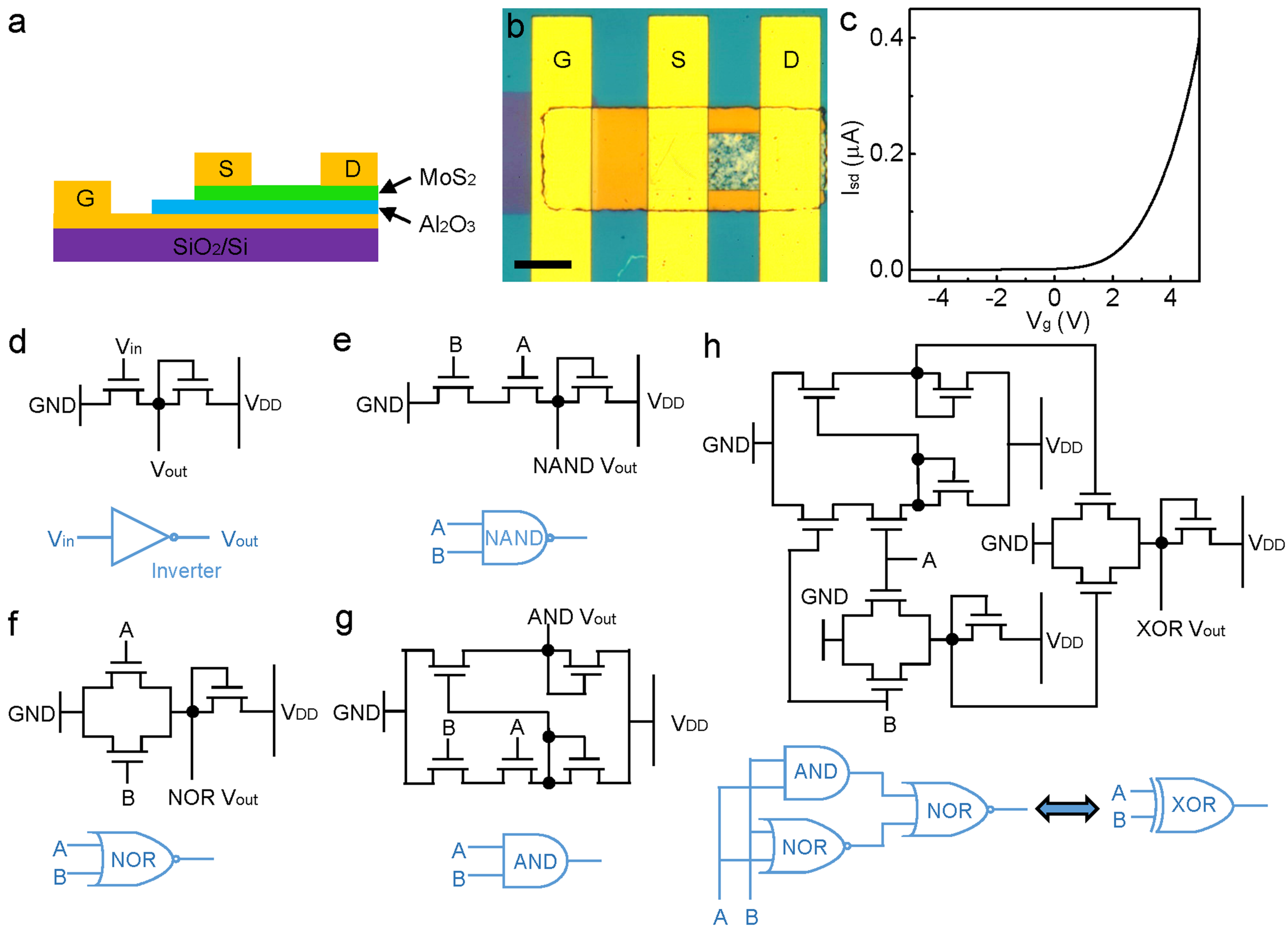
Extended Data Fig. 5 | Output characteristics of 50 independently measured transistors. V_g ranges from −10 V (black) to 40 V (dark blue) in steps of 10 V.



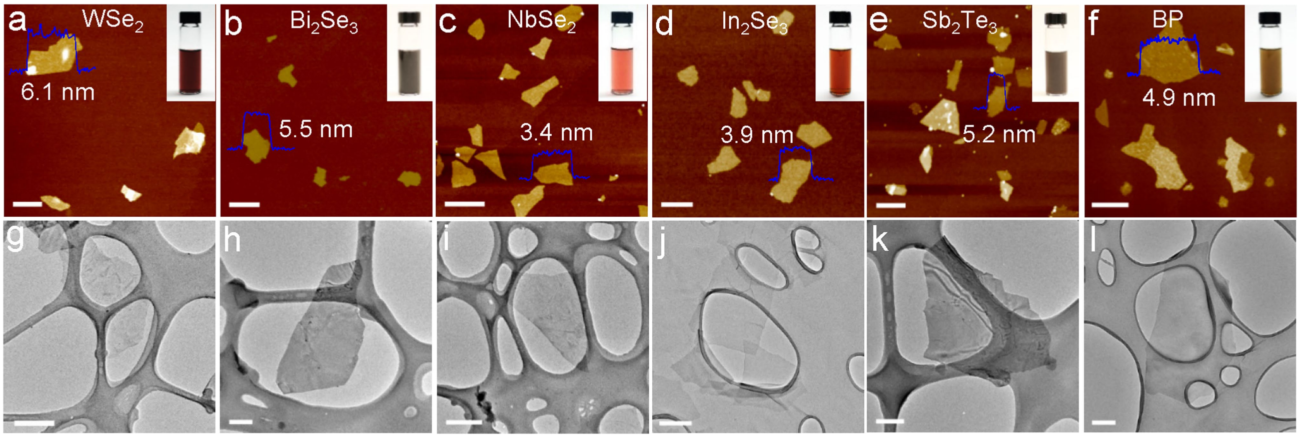
Extended Data Fig. 6 | Transfer characteristics of 50 independently measured transistors. a, **b**, I_{sd} on linear (**a**) and logarithmic (**b**) scales. $V_{sd} = 1$ V.



Extended Data Fig. 7 | I_{sd}–V_g transfer characteristics of MoS_2 films annealed at various temperatures. a, Annealed at 200 °C. **b**, Annealed at 400 °C. Both devices were fabricated on a 90-nm-thick SiO_2/Si substrate, with V_{sd} = 1 V. The threshold voltage shifts to more positive values when the film is annealed at a higher temperature, suggesting a reduced carrier concentration owing to the removal of impurity doping at the higher temperature.



Extended Data Fig. 8 | Constructing logic gates and computational circuits using MoS_2 thin-film transistors with a local back gate. **a**, Schematic of the MoS_2 transistor with a local back gate (Ti/Au, 10 nm/40 nm) buried under a 30-nm-thick Al_2O_3 dielectric layer. **b**, Optical image of an individual transistor. The gate electrode was fabricated after etching the underlying Al_2O_3 layer. Scale bar, 20 μm. **c**, I_{sd}–V_g transfer characteristics of the transistor constructed on the buried gate with a 30-nm-thick Al_2O_3 dielectric layer and $V_{sd} = 1$ V. The device was annealed at 400 °C for 1 h. **d**–**h**, Schematics of the logic inverter (**d**) and the NAND (**e**), NOR (**f**), AND (**g**) and XOR (**h**) gates. GND, ground.



Extended Data Fig. 9 | Exfoliation of various 2D crystals. **a**–**l**, AFM (**a**–**f**) and TEM (**g**–**l**) images of exfoliated nanosheets of WSe_2 (**a**, **g**), Bi_2Se_3 (**b**, **h**), $NbSe_2$ (**c**, **i**), In_2Se_3 (**d**, **j**), Sb_2Te_3 (**e**, **k**) and black phosphorus (BP; **f**, **l**). Insets are photographs of the corresponding ink solutions. Scale bars, 1 μm (**a**, **c**, **f**), 500 nm (**b**, **d**, **e**, **g**, **i**, **l**) and 200 nm (**h**, **j**, **k**).

Extended Data Table 1 | Comparison of device performance at room temperature for polycrystalline 2D semiconductor thin films

Material	Processing temperature (°C)	Mobility ($cm^2 \cdot V^{-1} \cdot s^{-1}$)	On/Off ratio	Reference
Solution-processed MoS_2 thin film	25	0.02	N/A	*(32)*
	50	< 0.3	< 2	*(12)*
	70	0.15	10^2	*(14)*
	300	7-11	10^6	*This work*
	350	0.4 (not pure MoS_2)	10^6	*(16)*
	450	< 0.1	< 10	*(33)*
CVD-grown MoS_2 thin film	450	14	10^2	*(27)*
	550	29 (high-*k* dielectric)	10^6	*(7)*
	650	17 (high-*k* dielectric)	10^6	*(26)*
	750	1.8	10^4	*(34)*
	750	5	10^5	*(35)*
	850	0.03	< 10^4	*(36)*
	850	16	10^6	*(37)*
	900	1.3	10^6	*(38)*
	1000	0.01	10^2	*(39)*
	1000	0.1	10^2	*(40)*
	1000	4.7	10^5	*(41)*
	700	45 (single crystal domain)	10^6	*(42)*
	900	67 (single crystal domain)	10^6	*(43)*
Solution-processed $MoSe_2$ film	70	0.18	< 10^2	*(14)*
Solution-processed WS_2 film	70	0.22	< 10^3	*(14)*
Solution-processed WSe_2 film	70	0.08	< 10^2	*(14)*

Data are from this work and refs [7,12,14,16,26,27,32–43]. N/A, not applicable; *k*, dielectric constant.

https://doi.org/10.1038/s41586-018-0577-1

Trade-offs in using European forests to meet climate objectives

Sebastiaan Luyssaert[1,2]*, Guillaume Marie[1], Aude Valade[3,5], Yi-Ying Chen[2,6], Sylvestre Njakou Djomo[4], James Ryder[2,7], Juliane Otto[2,8], Kim Naudts[2,9], Anne Sofie Lansø[2], Josefine Ghattas[3] & Matthew J. McGrath[2]

The Paris Agreement promotes forest management as a pathway towards halting climate warming through the reduction of carbon dioxide (CO_2) emissions[1]. However, the climate benefits from carbon sequestration through forest management may be reinforced, counteracted or even offset by concurrent management-induced changes in surface albedo, land-surface roughness, emissions of biogenic volatile organic compounds, transpiration and sensible heat flux[2–4]. Consequently, forest management could offset CO_2 emissions without halting global temperature rise. It therefore remains to be confirmed whether commonly proposed sustainable European forest-management portfolios would comply with the Paris Agreement—that is, whether they can reduce the growth rate of atmospheric CO_2, reduce the radiative imbalance at the top of the atmosphere, and neither increase the near-surface air temperature nor decrease precipitation by the end of the twenty-first century. Here we show that the portfolio made up of management systems that locally maximize the carbon sink through carbon sequestration, wood use and product and energy substitution reduces the growth rate of atmospheric CO_2, but does not meet any of the other criteria. The portfolios that maximize the carbon sink or forest albedo pass only one—different in each case—criterion. Managing the European forests with the objective of reducing near-surface air temperature, on the other hand, will also reduce the atmospheric CO_2 growth rate, thus meeting two of the four criteria. Trade-off are thus unavoidable when using European forests to meet climate objectives. Furthermore, our results demonstrate that if present-day forest cover is sustained, the additional climate benefits achieved through forest management would be modest and local, rather than global. On the basis of these findings, we argue that Europe should not rely on forest management to mitigate climate change. The modest climate effects from changes in forest management imply, however, that if adaptation to future climate were to require large-scale changes in species composition and silvicultural systems over Europe[5,6], the forests could be adapted to climate change with neither positive nor negative climate effects.

Following the Paris Agreement, the European Union and its 28 member states have committed to a 40% domestic reduction in greenhouse-gas emissions compared to 1990 levels by 2030. About 75% of this reduction is expected to come from emission reductions and the remaining 25% from land use, land-use change and forestry[7]. The commitment to reduce domestic greenhouse-gas emissions through forestry is in turn reflected in the national strategies of several European countries for energy, climate change and forestry[8–10]. These strategies typically focus on enhancing forestry-based sinks and reservoirs and developing neutral- or negative-emission approaches based on woody biomass. Furthermore, European forest owners who have reported to have experienced climate change have indicated that this experience influenced their management decisions[11]. Hence, climate change and the Paris Agreement are already shaping forest-management decisions. Despite being explicitly mentioned in both the Kyoto Protocol[12] and the Paris Agreement[1], little is known about the climate effects of forest management, including the effects of human-induced changes in tree species and silvicultural systems[3,13,14].

This study searches for spatially explicit forest-management portfolios for Europe that comply with the Paris Agreement up to the turn of the twenty-first century. The agreement requires that forest management jointly reduces the growth rate of atmospheric CO_2 (Articles 4 and 5) and the radiative imbalance at the top of the atmosphere (Article 2). Furthermore, forest management compliant with the Paris Agreement should neither increase the near-surface air temperature (hereafter referred to as 'air temperature') nor decrease precipitation, because changing the climate of the terrestrial biosphere would make adaptation to climate change (Article 7) even more difficult (see Supplementary Information, 'Operationalizing the Paris Agreement').

Simulation experiments that combine vegetation modelling, climate modelling, vegetation–climate feedbacks and life-cycle analysis are used to quantify the CO_2 emissions, radiative imbalance at the top of the atmosphere, air temperature and precipitation of three spatially explicit forest-management portfolios for Europe (Extended Data Fig. 1). Each portfolio has a distinct objective: maximize the forest carbon sink, maximize forest albedo or reduce air temperature.

All portfolios start from the same 2010 species and age–class distribution. Once an individual forest reaches maturity, six scenarios are explored: (i) refrain from harvesting; (ii) harvest, replant the same species and apply the same silvicultural system as before; (iii) harvest, replant the same species and thin before the final felling; (iv) harvest, change to the most common deciduous species in that region and thin before the final felling; (v) harvest, change to the most common deciduous species in that region and manage it as a coppice; and (vi) harvest, change to the most common conifer species in that region and thin before the final felling. Subsequently, portfolios are constructed by selecting the best-performing management scenario for each of the three objectives and for each $0.5° \times 0.5°$ grid cell in the European domain.

In contrast to previous land-use simulation experiments, our portfolios simulate a realistic rate of change for tree-species distributions and silvicultural systems because changes are only implemented following a harvest or stand-replacing mortality. Thus, management changes are dictated by forest growth and human choices within natural constraints, rather than through externally prescribed harvest volumes or through strictly natural succession.

A management portfolio that maximizes the carbon sink[15,16] reflects the widely held view that the net climate effect of forest management is dominated by decreasing the growth rate of atmospheric CO_2 through forest-based carbon sequestration, carbon storage in wood products, and material and energy substitution. Implementing the sink-maximizing portfolio—instead of the business-as-usual one—would require converting 475,000 km^2 of deciduous forest in central and southern Europe

[1]Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [2]Laboratoire des Sciences du Climat et de l'Environnement (LSCE/IPSL), CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France. [3]Institut Pierre Simon Laplace (IPSL), Paris, France. [4]Department of Agroecology, Aarhus University, Tjele, Denmark. [5]Present address: Global Ecology Unit CREAF-UAB, Cerdanyola del Vallès, Spain. [6]Present address: Research Center for Environmental Changes (RCEC), Academia Sinica, Taipei, Taiwan. [7]Present address: National Physical Laboratory, Teddington, London, UK. [8]Present address: Helmholtz-Zentrum Geesthacht (HZG), Climate Service Center Germany (GERICS), Hamburg, Germany. [9]Present address: Max Planck Institute for Meteorology, Hamburg, Germany. *e-mail: s.luyssaert@vu.nl

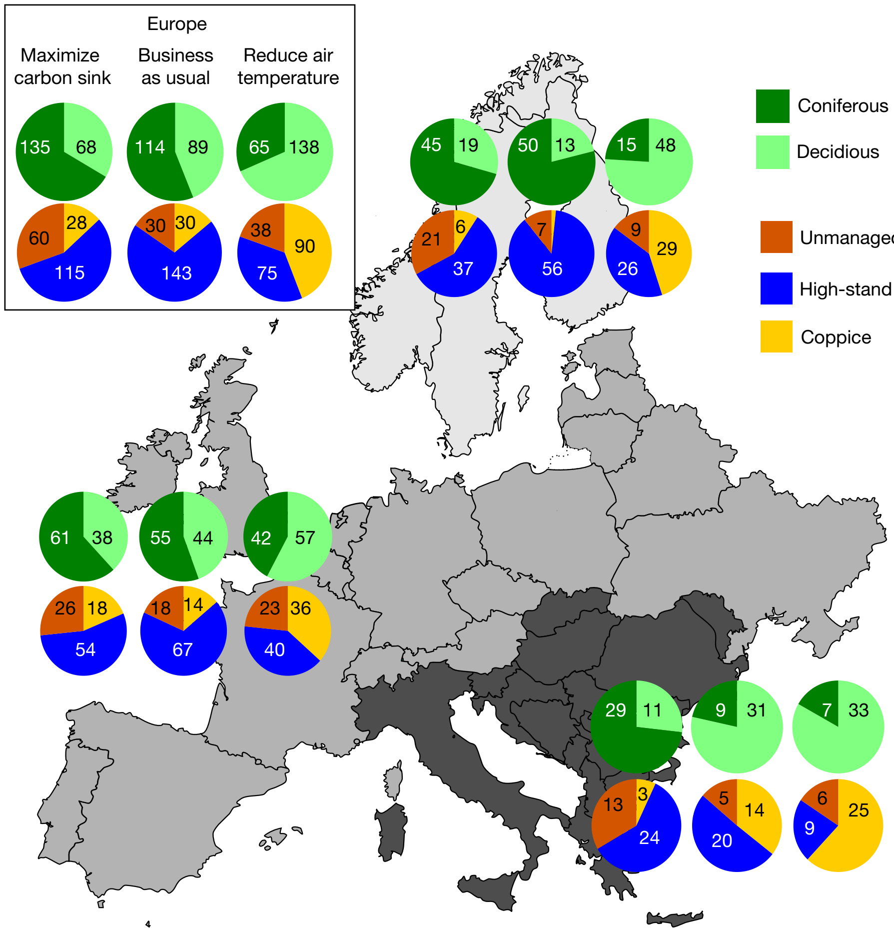


Fig. 1 | Forest surface areas (×10,000 km²) by 2100 under different forest-management portfolios. The portfolios considered here maximize the carbon sink, extend present-day management (business as usual) and reduce the air temperature. Forest management approaches include changes in tree species composition and silvicultural systems. The inset presents the mean values for all of Europe. Regional differences are shown for three geographical regions, each shown in a different shade of grey.

into coniferous forest, whereas 266,000 km^2 of previously coniferous forests in northern and central Europe would have to be converted to deciduous forests (Fig. 1; Extended Data Table 1; see Supplementary Information, 'Drivers of changes in forest management').

A sink-maximizing portfolio would come with a 12% lower wood harvest but could offset an additional 8.1 Pg C (1 Pg $= 10^{15}$ g) of fossil-fuel emissions (Table 1) between 2010 and 2100 compared with a business-as-usual management portfolio, which extends the present-day forest-management portfolio into the future. This increase in the projected carbon savings is similar to estimates reported by the forestry sector[16] and could be achieved by optimizing the balance between forest-based sequestration (8.2 Pg C) on the one hand and product-based sinks and substitution (−0.3 Pg C), energy-based substitution (0.2 Pg C) and savings in the emissions from forest exploitation, wood processing and product manufacturing (0.05 Pg C) on the other. Accounting for ocean uptake of atmospheric CO_2 (see Supplementary Information, 'Life cycle analysis') results in a cumulated net reduction of the atmospheric CO_2 concentration of 4.3 Pg C in 2100, which translates into a 2 p.p.m. decrease in atmospheric CO_2 compared with the business-as-usual portfolio (Table 1). Owing to the changes in tree species and silvicultural systems that are required to realize this 2 p.p.m. reduction, the approximately 0.002 W m^{-2} decrease in the radiative imbalance at the top of the atmosphere from the stronger carbon sink[17] is neutralized by unintended, but unavoidable, changes in surface albedo (−0.001) and cloud cover (−0.1%). The carbon-sink-maximizing portfolio has a small negative effect on annual (−2 mm) and no effect on air temperature (Table 1).

A temperature-based portfolio reflects the idea that management-induced changes in surface properties may redistribute heat away from the land surface, resulting in a local cooling of the land surface[18] that can be beneficial for organisms living there. Implementing such a portfolio requires converting 493,000 km^2 of coniferous forests to deciduous forests (of which 65% would be in Scandinavia) and coppicing an additional 600,000 km^2 of deciduous forests (Fig. 1; Extended Data Table 1; see Supplementary Information, 'Drivers of changes in forest management'). Such changes in forest management would, however, reduce the wood harvest by 25% compared to the business-as-usual portfolio (Table 1). By 2100 these changes would result in a cumulative net reduction of the atmospheric CO_2 concentration of 1.8 Pg C, which is equivalent to a 0.9 p.p.m. reduction of atmospheric CO_2 compared with the business-as-usual portfolio (Table 1).

The combined biogeochemical and biophysical effects of this portfolio come without a significant effect on the radiative imbalance at the top of the atmosphere (one-sided t-test, $P = 0.28$), but could contribute to a 0.3 K cooling over Scandinavia, with a much smaller effect on temperature over the rest of Europe (Fig. 2a). Following a large-scale transition to deciduous species, cooling of the air temperature is projected to occur only in winter and spring (Extended Data Fig. 2). In spring, air-temperature cooling from an increase in surface albedo due to decreased snow masking by deciduous canopies would be partly compensated by warming from a decrease in turbulent fluxes caused by the absence of leaves until bud break later in spring (Fig. 2b). Our simulation experiment thus confirms the role of transpiration in determining air temperature, even at high latitudes[19].

A portfolio that maximizes the albedo[20] reflects the view that managing the forest albedo would reduce the radiative imbalance at the top of the atmosphere while maintaining the forest carbon sink.

Table 1 | Biogeochemical and biophysical effects over Europe in 2100 for four different forest-management portfolios

Variable (units)	Business as usual	Maximize carbon sink	Maximize albedo	Reduce air temperature
Global average TOA (W m^{-2})	4.31 ± 0.01	4.31	4.33	4.32
Change in CO_2 sink and avoided emissions between 2010 and 2100 (Pg C)	4.7	12.8	5.0	8.1
Change in net cumulated atmospheric CO_2 between 2010 and 2100 (Pg C)	−2.7	−7.0	−2.8	−4.5
Atmospheric CO_2 (p.p.m.)	934.6	932.6	934.6	933.8
Air temperature (K)	283.84 ± 0.001[a]	283.84	283.83	283.81
Annual precipitation (mm)	734.7 ± 0.1	732.6	730.0	730.9
Summer precipitation (mm)	166.1 ± 0.1	165.2	163.7	165.0
Wood harvest (Tg C y^{-1})	203.2	179.5	144.5	151.6
Surface albedo (–)	0.113 ± 0.0001[a]	0.113	0.128	0.126
Evapotranspiration (mm)	555.5 ± 0.1	552.8	546.4	549.2
Latent heat (W m^{-2})	44.35 ± 0.01[a]	44.13	43.60	43.82
Sensible heat (W m^{-2})	26.67 ± 0.01[a]	26.82	27.28	27.00
Total cloud cover (%)	46.8 ± 0.1[a]	46.7	46.7	46.6

The business-as-usual simulation, which served as a control, was repeated three times with slightly different initial atmospheric conditions (see Supplementary Information, 'Equilibrium climate for the management portfolios'). The variability between these three repetitions was considered to be the minimal model noise of the climate model and to define one standard deviation. TOA denotes the radiative imbalance at the top of the atmosphere. Results for two additional portfolios are presented in Extended Data Table 2.
[a]Upper limit.

Our simulations confirm that an albedo-maximizing portfolio would decrease wood harvest by 30% and realize cumulated net emission savings of up to 2.8 Pg C, which is comparable to the savings expected from the business-as-usual portfolio. However, the increase in surface albedo that can be realized through the albedo-based portfolio (+0.015) would be compensated by a decrease in cloud cover (−0.1%) and therefore come without a significant effect on the radiative imbalance at the top of the atmosphere (one-sided t-test, $P = 0.07$) and with a small negative effect on air temperature (−0.01 K; Table 1).

Furthermore, all portfolios reduce the mean annual precipitation by 2.1–4.7 mm compared to the business-as-usual portfolio. Reductions are evenly spread across the seasons and consistent with the decrease in cloud cover and evapotranspiration (Table 1). Hence, none of the tested forest-management portfolios meets all of the four criteria set for compliance with the Paris Agreement. Maximizing the carbon sink and maximizing the forest albedo both meet one of the four criteria. Managing European forests with the objective of reducing air temperature satisfies two of the four criteria: reducing the air temperature and reducing the CO_2 growth rate. Therefore, making trade-offs seems unavoidable when using European forests to meet climate objectives.

To our knowledge, this study is the first to quantify the capacity of forest management to comply with the Paris Agreement while addressing both biogeochemical and biophysical effects; hence, its results could not be compared with previous reports. The small temperature effects, compared with those found in global afforestation and deforestation studies[21–24], are thought to be the consequence of considering a realistic 90-year period of management changes and testing the portfolios for a limited global land area, that is, about 7% of the global total of managed forest[14]. Although a global implementation of carbon-based forest management will probably enhance the carbon sink of the forest sector globally[15], the combined biogeochemical and biophysical effects cannot be extrapolated from Europe to the global scale owing to biome-specific land–atmosphere interactions[25,26].

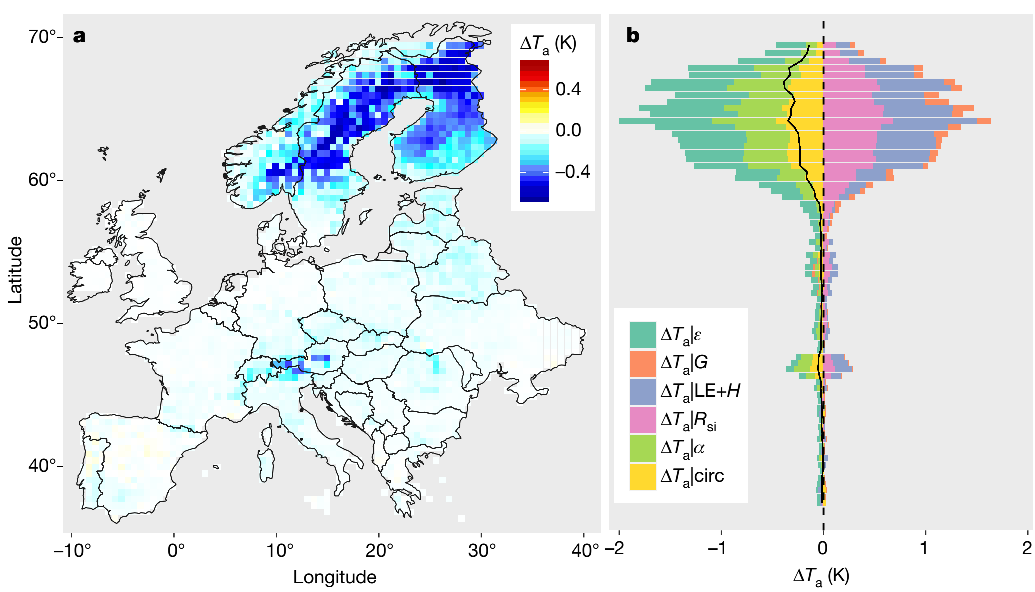


Fig. 2 | Changes and main drivers of air temperature in February and March by the turn of the 21st century for a forest-management portfolio that reduces the near-surface air temperature. **a**, Spatially explicit changes in air temperature (ΔT_a) in February and March. Temperature changes smaller than 1.96σ are shown in white, where the standard deviation σ represents the minimal noise of the simulation code LMDzORCAN (see Supplementary Information, 'Equilibrium climate for the management portfolios'). **b**, Drivers of the changes in springtime air temperature for 0.5° latitudinal bands. Shown are air temperature changes due to changes in atmospheric emissivity ($\Delta T_a|\varepsilon$), ground heat flux ($\Delta T_a|G$), turbulent fluxes ($\Delta T_a|\mathrm{LE}+H$), shortwave incoming radiation ($\Delta T_a|R_{si}$), which in this simulation experiment is a proxy for cloud cover, surface albedo ($\Delta T_a|\alpha$) and atmospheric circulation ($\Delta T_a|\mathrm{circ}$). See Supplementary Information for details.

A global implementation of locally optimized forest-management portfolios would lead to larger areas with near-surface cooling. Given that air temperature cooling was found to saturate quickly with the fractional change in tree species composition (Extended Data Fig. 3), the magnitude of the cooling is not expected to change substantially following a large-scale implementation, unless ocean feedbacks[19,22], cloud feedbacks through species-specific biogenic emissions of volatile organic compounds[27], and changes in the North Atlantic Oscillation[28], which were not fully accounted for in this study, are among the key drivers.

Our results demonstrate that, on the basis of a single model, in the absence of carbon capture and storage the additional climate benefits of sustainable forest management will be modest and local rather than global. Hence, we suggest that the primary role of forest management in Europe in the coming decades is not to protect the climate, but to adapt the forest cover to future climate[5] in order to sustain the provision of wood and ecological, social and cultural services[29], while avoiding positive climate feedbacks from fire, wind, pests and drought disturbances[30]. Even if adaptation would require large-scale changes in the tree species composition and silvicultural systems over Europe[5,6], our results imply that these changes themselves will probably have little impact on the climate.

Code availability

The code and the run environment used in this study are open-source and distributed under the CeCILL (CEA CNRS INRIA Logiciel Libre) license. The codes of ORCHIDEE-CAN_r2290 and ORCHIDEE-CAN_r3069 can be accessed at https://doi.org/10.14768/06337394-73A9-407C-9997-0E380DAC5595 and https://doi.org/10.14768/06337394-73A9-407C-9997-0E380DAC5596, respectively. Access to the run environment and LMDzORCAN are restricted to registered users; requests can be sent to the corresponding author. The post-processing code used to estimate the life-cycle sinks and emissions of the forestry sector (see Supplementary Information, 'Life cycle analysis'), search for the optimized management portfolios (see Supplementary Information, 'Management optimization' and decompose the air temperature into its main drivers (see Supplementary Information, 'Decomposing near-surface air temperature') can be accessed at https://doi.org/10.5281/zenodo.1284533.

Data availability

Figures 1, 2, Table 1, Extended Data Figs. 2, 3, Supplementary Fig. 1 and Extended Data Table 1, 2 are based on a simulation experiment whose output files (about 7.4 Tb) will be provided upon reasonable request. The data files that were used to set the boundary conditions of ORCHIDEE-CAN and LMDzORCAN (about 70 Gb) will be provided upon reasonable request.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0577-1.

Received: 29 March 2018; Accepted: 13 August 2018;
Published online 10 October 2018.

1. *Paris Agreement, FCCC/CP/2015/10/Add.1* https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf (United Nations Framework Convention on Climate Change, 2015).
2. Pielke, R. A. et al. Land use/land cover changes and climate: modeling analysis and observational evidence. *WIREs Clim. Change* **2**, 828–850 (2011).
3. Naudts, K. et al. Europe's forest management did not mitigate climate warming. *Science* **351**, 597–600 (2016).
4. Jackson, R. B. Trading water for carbon with biological carbon sequestration. *Science* **310**, 1944–1947 (2005).
5. Lindner, M. et al. Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. *For. Ecol. Manage.* **259**, 698–709 (2010).
6. Hanewinkel, M., Cullmann, D. A., Schelhaas, M.-J., Nabuurs, G.-J. & Zimmermann, N. E. Climate change may cause severe loss in the economic value of European forest land. *Nat. Clim. Chang.* **3**, 203–207 (2013).
7. Grassi, G. et al. The key role of forests in meeting climate targets requires science for credible mitigation. *Nat. Clim. Chang.* **7**, 220–226 (2017).
8. *National Action Plans* https://ec.europa.eu/energy/en/topics/renewable-energy/national-action-plans (European Commission, 2017).
9. *Government Report on the National Energy and Climate Strategy for 2030* http://julkaisut.valtioneuvosto.fi/handle/10024/79247 (Finnish Ministry of Economic Affairs and Employment, 2017).
10. *Décret N° 2017–155 du 8 Février 2017 Portant Approbation du Programme National de la Forêt et du Bois* https://www.legifrance.gouv.fr/eli/decret/2017/2/8/AGRT1632600D/jo/texte (Legifrance, 2016).
11. Blennow, K., Persson, J., Tomé, M. & Hanewinkel, M. Climate change: believing and seeing implies adapting. *PLoS One* **7**, e50182 (2012).
12. *Kyoto Protocol to the United Nations Framework Convention on Climate Change* https://unfccc.int/sites/default/files/kpeng.pdf (United Nations, 1998).
13. Erb, K.-H. et al. Land management: data availability and process understanding for global change studies. *Glob. Change Biol.* **23**, 512–533 (2017).
14. Luyssaert, S. et al. Land management and land-cover change have impacts of similar magnitude on surface temperature. *Nat. Clim. Chang.* **4**, 389–393 (2014).
15. Canadell, J. G. & Raupach, M. R. Managing forests for climate change mitigation. *Science* **320**, 1456–1457 (2008).
16. Nabuurs, G. et al. *A New Role for Forests and the Forest Sector in the EU Post-2020 Climate Targets* https://www.efi.int/sites/default/files/files/publication-bank/2018/efi_fstp_2_2015.pdf (European Forest Institute, 2015).
17. Myhre, G., Highwood, E. J., Shine, K. P. & Stordal, F. New estimates of radiative forcing due to well mixed greenhouse gases. *Geophys. Res. Lett.* **25**, 2715–2718 (1998).
18. Davin, E. L., Seneviratne, S. I., Ciais, P., Olioso, A. & Wang, T. Preferential cooling of hot extremes from cropland albedo management. *Proc. Natl Acad. Sci. USA* **111**, 9757–9761 (2014).
19. Swann, A. L., Fung, I. Y., Levis, S., Bonan, G. B. & Doney, S. C. Changes in Arctic vegetation amplify high-latitude warming through the greenhouse effect. *Proc. Natl Acad. Sci. USA* **107**, 1295–1300 (2010).
20. Lutz, D. A. & Howarth, R. B. Valuing albedo as an ecosystem service: Implications for forest management. *Clim. Change* **124**, 53–63 (2014).
21. Arora, V. K. & Montenegro, A. Small temperature benefits provided by realistic afforestation efforts. *Nat. Geosci.* **4**, 514–518 (2011).
22. Davin, E. L. & de Noblet-Ducoudré, N. Climatic impact of global-scale deforestation: radiative versus nonradiative processes. *J. Clim.* **23**, 97–112 (2010).
23. Bathiany, S., Claussen, M., Brovkin, V., Raddatz, T. & Gayler, V. Combined biogeophysical and biogeochemical effects of large-scale forest cover changes in the MPI earth system model. *Biogeosciences* **7**, 1383–1399 (2010).
24. Bala, G. et al. Combined climate and carbon-cycle effects of large-scale deforestation. *Proc. Natl Acad. Sci. USA* **104**, 6550–6555 (2007); correction **104**, 9911 (2007).
25. Beringer, J., Chapin, F. S., Thompson, C. C. & McGuire, A. D. Surface energy exchanges along a tundra-forest transition and feedbacks to climate. *Agric. For. Meteorol.* **131**, 143–161 (2005).
26. Alkama, R. & Cescatti, A. Biophysical climate impacts of recent changes in global forest cover. *Science* **351**, 600–604 (2016).
27. Unger, N. Human land-use-driven reduction of forest volatiles cools global climate. *Nat. Clim. Chang.* **4**, 907–910 (2014).
28. Hurrell, J. W. Decadal trends in the North Atlantic oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
29. Costanza, R. et al. The value of the world's ecosystem services and natural capital. *Nature* **387**, 253–260 (1997).
30. Seidl, R., Schelhaas, M.-J., Rammer, W. & Verkerk, P. J. Increasing forest disturbances in Europe and their impact on carbon storage. *Nat. Clim. Chang.* **4**, 806–810 (2014); corrigendum **4**, 930 (2014).

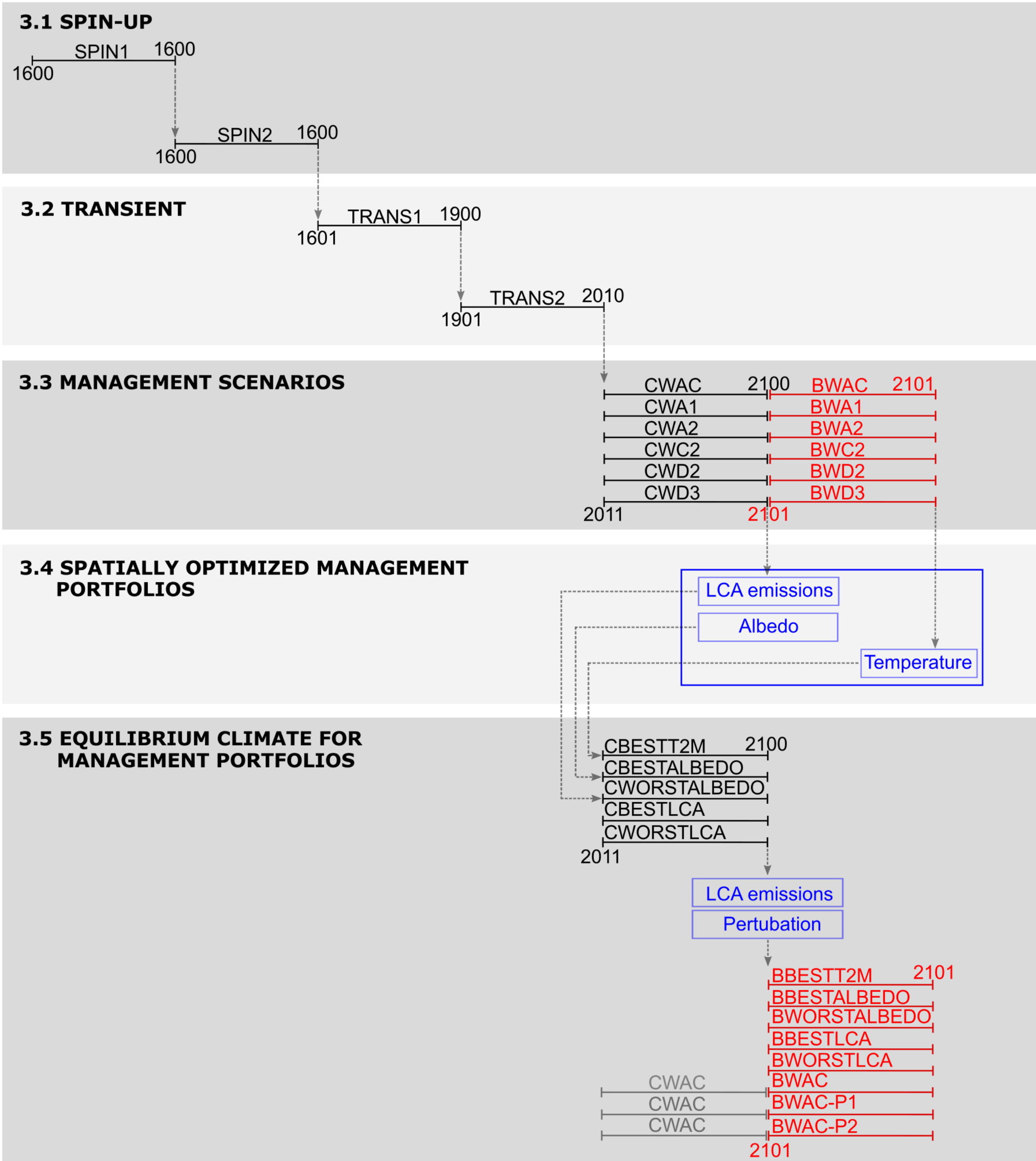
Acknowledgements M.J.M., K.N., J.R., Y.-Y.C., J.O. and S.L. were funded through the European Research Council (ERC) starting grant 242564 and A.V. was funded through the Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME). S.L. and G.M. were partly funded through an Amsterdam Academic Alliance (AAA) fellowship. S.L. is grateful for the mentorship of E.-D. Schulze, I. A. Janssens and P. Ciais. The ORCHIDEE and LMDZ project teams and the Centre de Calcul Recherche et Technologie (CCRT) provided the run environment that enabled the land–atmosphere simulations conducted in this study.

Reviewer information *Nature* thanks T. Pugh and K. Zhao for their contribution to the peer review of this work.

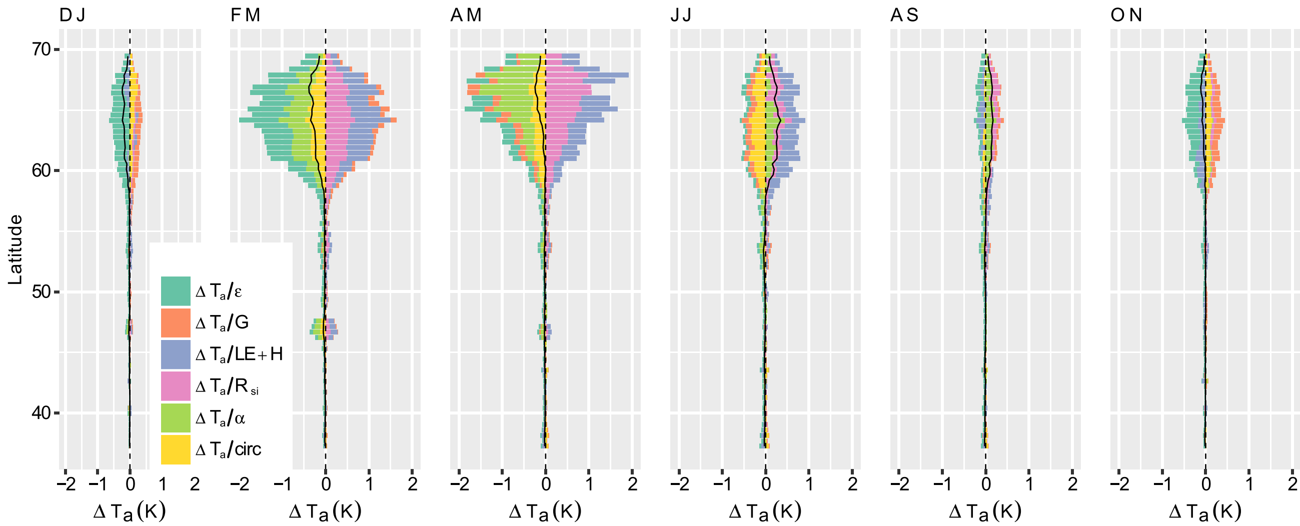
Author contributions S.L. and M.J.M. designed the study. M.J.M., J.O., J.R., Y.-Y.C., K.N., A.V. and S.L. developed, parameterized and validated ORCHIDEE-CAN. G.M., M.J.M., J.G. and S.L. conducted the simulation experiment. S.N.D. developed the life-cycle analysis method. G.M., Y.-Y.C. and S.L. analysed the data. G.M., M.J.M., J.O., J.R., Y.-Y.C., K.N., A.V., A.S.L. and S.L. contributed to the interpretation of the results.

Competing interests The authors declare no competing interests.

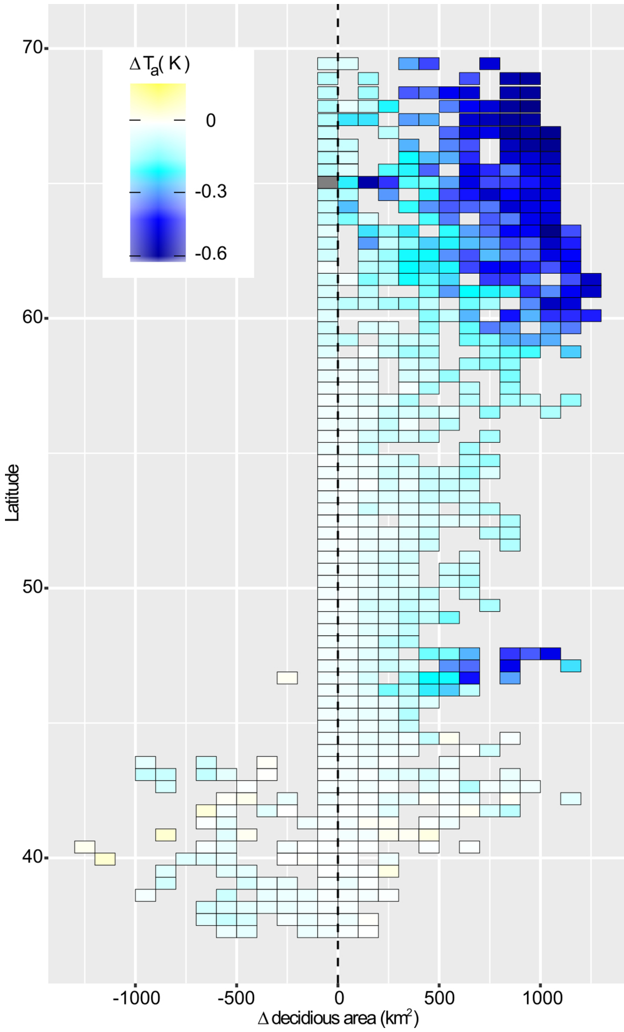
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0577-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0577-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Setup of the simulation experiments. The experiments are described in the section 'Simulation experiment' in Supplementary Information. Simulations with ORCHIDEE-CAN are shown in black and simulations with LMDzORCAN are shown in red. Blue boxes denote intermediate calculations using the simulation results (see Supplementary Information, 'Spatially optimized management portfolios' and 'Equilibrium climate for the management portfolios'). The simulations shown in this figure correspond to runs with reduced air temperature (BBESTT2M), maximized surface albedo (BESTALBEDO), minimized surface albedo (BWORSTALBEDO), maximized carbon sink (BBESTLCA), minimized carbon sink (BWORSTLCA) and business as usual (BWAC). BWAC, BWAC-P1 and BWAC-P2 were used to calculate the minimal model noise.



Extended Data Fig. 2 | Drivers of the mean bimonthly air temperature changes for 0.5° latitudinal bands. The notation is as in Fig. 2 and the labels at the top denote months (D J, December and January; F M, February and March; A M, April and May; and so on). Although all the components contribute to the change of the air temperature, changes in emissivity always result in cooling and changes in shortwave incoming radiation always result in warming. Consequently, emissivity and incoming shortwave radiation cannot explain the seasonal variation in air temperature changes. The other components are positively correlated with air temperature in some months and negatively correlated in others, which rules them out as the main driver of air temperature changes and suggests that the net effect is the outcome of the interplay between the different components.



Extended Data Fig. 3 | Relationship between changes in springtime air temperature and changes in the fractional cover of deciduous forest for 0.5° latitudinal bands over Europe. Locations where the tree species are maintained between 2010 and 2100 (that is, the difference Δ of the deciduous area is 0) could experience similar air temperature changes as neighbouring locations where one tree species is replaced by another, especially in Scandinavia, suggesting advection of heat and moisture. Nevertheless, at lower latitudes the spatial scale of this advection is limited to a few pixels (for example, Fig. 2a) corresponding to a range of 50–200 km. Furthermore, the temperature effect quickly saturates with the fractional cover change and shows a strong dependence on geographical location (see Supplementary Information). Whether this apparent geographical dependence is the outcome of climatic differences or of differences between northern and southern European deciduous species could not be established with the experimental setup used in this study.

Extended Data Table 1 | Changes in surface area of European forests by 2100 for six different forest-management portfolios

Change in surface area (km^2)	Business as usual (BAU)	Maximise carbon sink	Maximise albedo	Minimise carbon sink	Minimise albedo	Reduce near-surface temperature
Deciduous to conifers	0	475,000	30,000	6,000	516,000	41,000
Conifers to deciduous	0	266,000	590,000	236,000	26,000	534,000
Net increase conifers	0	209,000	-560,000	-230,000	490,000	-493,000
Net increase thin and fell	0	-280,000	-330,000	-390,000	-230,000	-680,000
Net increase coppice	0	-20,000	130,000	-130,000	-210,000	600,000
Net increase unmanaged	0	300,000	200,000	520,000	440,000	80,000

We note that the total surface area of forests was held constant at 2,000,000 km^2 between 2010 and 2100 for reasons described in Supplementary Information, 'Simulation experiment'.

Extended Data Table 2 | Biogeochemical and biophysical effects over Europe in 2100 for two forest-management portfolios

Variable name (units)	Minimise carbon sink	Minimise albedo
TOA (W m^{-2})	4.32	4.32
Change in CO_2 sink & avoided emissions between 2010 and 2100(Pg C)	0.7	10.5
Change in net cumulated atmospheric CO_2 between 2010 and 2100 (Pg C)	- 0.5	- 5.7
Atmospheric CO_2 (ppm)	935.7	933.2
Near surface temperature (K)	283.85	283.86
Annual precipitation (mm)	733.1	734.2
Summer precipitation (mm)	164.0	165.4
Wood harvest (Tg C y^{-1})	122.9	176.2
Surface albedo (-)	0.119	0.107
Evapotranspiration (mm)	550.0	553.9
Latent heat (W m^{-2})	43.90	44.23
Sensible heat (W m^{-2})	27.12	26.81
Total cloud cover (%)	46.8	46.8

https://doi.org/10.1038/s41586-018-0582-4

Effects of climate warming on photosynthesis in boreal tree species depend on soil moisture

Peter B. Reich[1,2]*, Kerrie M. Sendall[1,3], Artur Stefanski[1], Roy L. Rich[1,4], Sarah E. Hobbie[5] & Rebecca A. Montgomery[1]

Climate warming will influence photosynthesis via thermal effects and by altering soil moisture[1–11]. Both effects may be important for the vast areas of global forests that fluctuate between periods when cool temperatures limit photosynthesis and periods when soil moisture may be limiting to carbon gain[4–6,9–11]. Here we show that the effects of climate warming flip from positive to negative as southern boreal forests transition from rainy to modestly dry periods during the growing season. In a three-year open-air warming experiment with juveniles of 11 temperate and boreal tree species, an increase of 3.4 °C in temperature increased light-saturated net photosynthesis and leaf diffusive conductance on average on the one-third of days with the wettest soils. In all 11 species, leaf diffusive conductance and, as a result, light-saturated net photosynthesis decreased during dry spells, and did so more sharply in warmed plants than in plants at ambient temperatures. Consequently, across the 11 species, warming reduced light-saturated net photosynthesis on the two-thirds of days with driest soils. Thus, low soil moisture may reduce, or even reverse, the potential benefits of climate warming on photosynthesis in mesic, seasonally cold environments, both during drought and in regularly occurring, modestly dry periods during the growing season.

A changing climate will influence plants by altering temperature, precipitation and soil moisture, as well as their variability and seasonality[1–11]. In temperate and boreal climates, temperatures switch seasonally from cold (and limiting to biological processes) to warm and periodically dry, during which time moisture can be limiting[2–6,9–11]. Both the 'law of the minimum' and multiple limitation theory[12–14] provide a conceptual basis for predicting climate warming interactions with soil moisture. Although higher temperatures may alleviate enzymatic limits on the biochemistry of photosynthesis, realized rates of CO_2 assimilation may decrease if and when low soil water causes stomatal closure and limitation of the CO_2 substrate for photosynthesis. As growing season conditions in temperate and boreal forests are likely to become effectively drier than in the past[3,8,9], because climate warming will increase evapotranspiration more than precipitation[3,9] and increase variability in the amount of precipitation per event[1,9], the importance of water availability to forest responses to rising temperature may increase in the future[3–6,9–11,15–18].

Mid- and high-latitude plants will therefore probably experience both positive and negative effects of climate warming on photosynthesis within and across years—we propose that these will be positive when soil moisture is ample but negative when soils are drier[4–6,9–11,15–17]. Whether such effects are in aggregate positive or negative is likely to depend on the balance of time that warming alleviates low temperature limitations to plant function as opposed to causing limitations to function through decreased soil moisture. However, direct tests of the effects of climate warming across a range of soil moisture conditions, caused by seasonal or interannual variation or by manipulations of temperature or moisture, are rare, and it remains unclear how plant responses to climate warming will be influenced by these indirect effects of soil moisture[4–6,9–11,16–18].

Here we provide evidence from 11 co-occurring boreal and temperate tree species (Fig. 1) in support of the overarching hypothesis that low soil moisture status has a dampening effect on the photosynthetic enhancement that results from experimental warming. This moisture regulation of the response to climate warming was consistent for all 11 species and occurred in response to reductions in soil moisture due to typical seasonal variation and in response to further reductions in soil moisture due to experimental warming. Results are from the free-air B4WarmED experiment[19–22], in which juveniles (3–5 years old at the time of measurements) of local ecotypes of the 11 tree species were grown under ambient and seasonally elevated (+3.4 °C, April–November) temperatures from 2009 to 2011 at two southern boreal sites in Minnesota, USA (Extended Data Table 1 and Methods). The 11 species co-occur in forests in northern Minnesota; however, five are boreal with southern range limits in or near Minnesota and six are temperate with northern range limits not far north of the Minnesota–Canada border[19]. Fluctuations in soil moisture levels (volumetric water content (VWC), m^3 H_2O per m^3 soil) occurred at both sites and across all years (Extended Data Fig. 1 and Extended Data Table 2), and spanned from 0.27 to 0.05 VWC, representing a range from slightly wetter than field capacity to slightly drier than the permanent wilting point (of approximately −1.5 MPa) for these sandy loam soils[23,24]. Leaf temperature (T_{leaf}) and vapour pressure gradient (VPG) also varied considerably across all photosynthetic measurements (Extended Data Fig. 2).

All species responses were consistent with the hypothesis that effects of experimental warming on carbon gain would be less positive or more negative during periods of low soil moisture (Fig. 1, Table 1 and Extended Data Table 3). In moist soils, all angiosperm species (and no gymnosperms) showed higher maximum carboxylation capacity at 25 °C ($V_{cmax\text{-}25}$) when grown at increased temperature compared to ambient temperatures (Extended Data Fig. 3), helping to explain the higher light-saturated net photosynthesis (A_{net}) in warmed plants when soil water limitations were modest (Fig. 1). This higher maximum carboxylation capacity in well-watered, warmed angiosperms assessed at a standardized temperature is indicative of an acclimation response (upregulation of $V_{cmax\text{-}25}$) to growth in elevated temperatures. However, every species showed marked sensitivity of A_{net} to drying soil moisture (Fig. 1). More relevant to our overarching hypothesis, A_{net} in all species declined more steeply with decreasing soil moisture in warmed than in ambient conditions (Fig. 1); therefore, when compared at a common soil moisture, plants showed the most positive (or least negative) effects of experimental warming on A_{net} when soil moisture availability was high, whereas positive effects decreased (or negative effects increased) as soil moisture availability declined (Fig. 1).

In other words, we found a significant interaction between the increased temperature treatment and VWC for A_{net} (Table 1; $F_{1,553} = 40.9$, $P < 0.0001$) in a model that included treatment (increased or ambient temperature), species, VWC and two other environmental drivers (T_{leaf} and VPG). Moreover, although species differed from each other in A_{net}, they did not differ in how VWC influenced their response

[1]Department of Forest Resources, University of Minnesota, St. Paul, MN, USA. [2]Hawkesbury Institute for the Environment, Western Sydney University, Penrith, New South Wales, Australia. [3]Department of Biology, Georgia Southern University, Statesboro, GA, USA. [4]Smithsonian Environmental Research Center, Edgewater, MD, USA. [5]Department of Ecology, Evolution and Behavior, University of Minnesota, St. Paul, MN, USA. *e-mail: preich@umn.edu

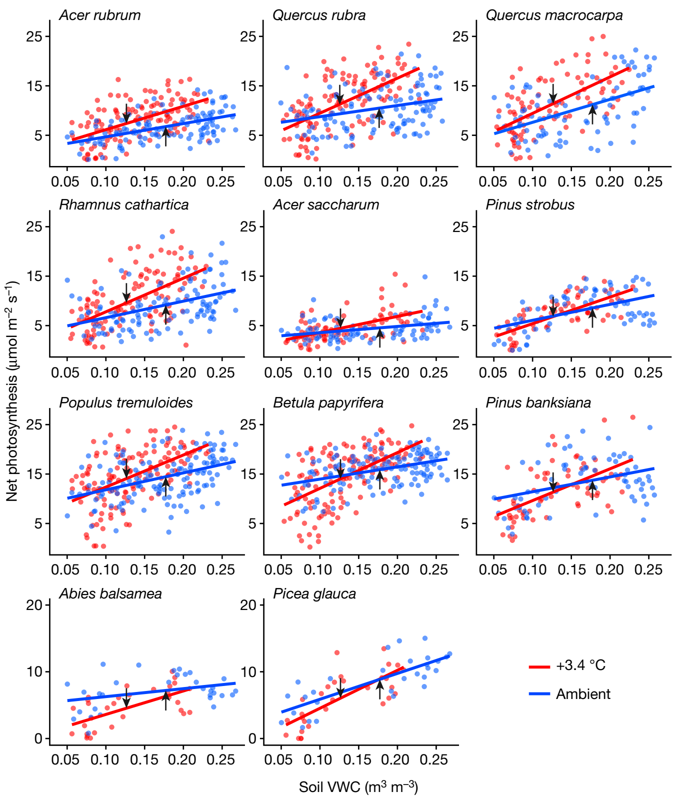


Fig. 1 | Photosynthesis is reduced by drying soils, and more so with simulated climate warming. In situ A_{net} in relation to soil moisture (VWC) by species for ambient (blue) and experimentally warmed (red) plants. Data are from multiple days across three years ($n = 1{,}991$ across species). The slope of A_{net} versus VWC was significantly steeper for warmed than for ambient plants (Table 1; $F_{1,553} = 40.9$, $P < 0.0001$). The arrows show the median VWC across all measurements for the ambient (up arrow) and warmed (down arrow) plants of each species. Species are arranged from top to bottom by their geographical ranges (temperate species in top two rows, boreal in bottom two rows). Sample sizes per species shown in Extended Data Table 3.

Table 1 | Summary of models of treatment and environmental effects on leaf gas exchange

	(A_{net})		(g_s)	
Source of variance	*F*	*P* > *F*	*F*	*P* > *F*
Species	**72.61**	**<0.0001**	**32.18**	**<0.0001**
Warm	**14.10**	**0.0003**	1.28	0.2587
Species × warm	**3.29**	**0.0003**	0.79	0.6430
Soil water	**215.61**	**<0.0001**	**147.72**	**<0.0001**
Soil water × species	**2.02**	**0.0278**	**6.17**	**<0.0001**
Soil water × warm	**40.88**	**<0.0001**	**6.44**	**0.0113**
Soil water × species × warm	1.17	0.3033	0.47	0.9130
VPG	**29.38**	**<0.0001**	**17.10**	**<0.0001**
VPG × species	**10.11**	**<0.0001**	**8.57**	**<0.0001**
VPG × warm	0.33	0.5686	0.42	0.5208
VPG × soil water	**5.59**	**0.0182**	0.30	0.5858
VPG × species × warm	1.39	0.1780	0.57	0.8427
VPG × species × soil water	**4.17**	**<0.0001**	1.35	0.1969
VPG × warm × soil water	**4.24**	**0.0396**	0.03	0.8629
T_{leaf}	**26.75**	**<0.0001**	3.32	0.0684
T_{leaf} × species	**11.77**	**<0.0001**	**6.65**	**<0.0001**
T_{leaf} × warm	0.05	0.8151	0.40	0.5251
T_{leaf} × soil water	**3.95**	**0.0469**	0.60	0.4382
T_{leaf} × VPG	0.69	0.4066	0.01	0.9157
T_{leaf} × species × warm	1.53	0.1225	0.55	0.8551
T_{leaf} × species × soil water	**3.46**	**0.0002**	1.59	0.1035
T_{leaf} × species × VPG	**2.39**	**0.0081**	1.70	0.0758
T_{leaf} × warm × soil water	**5.19**	**0.0228**	0.01	0.9047
T_{leaf} × warm × VPG	**3.46**	**0.0002**	0.01	0.9157
T_{leaf} × soil water × VPG	1.83	0.0502	0.19	0.6649
Full-model adjusted R^2	0.6342		0.6013	

Mixed models are shown for A_{net} and g_s in relation to species, +3.4 °C warming treatment (warm), volumetric water content (soil water), vapour pressure gradient (VPG), leaf temperature (T_{leaf}) and all interactions except the five-way interaction. Plot, block and site were included as random effects in the model. Both models were significant, at $P < 0.0001$. Data are for 11 species ($n = 1{,}991$ for A_{net}; 1,903 for g_s). Bold values indicate variables that are significant at $P < 0.05$. Four-way interactions were not significant and are not shown. *F* and *P* indicate *F*-statistics and *P* values, respectively.

to warming (no warming × soil moisture × species interaction, Table 1; $F_{10,1,797} = 1.2$, $P = 0.30$). Thus, species for which growth was enhanced (for example, *Acer* and *Quercus*) or reduced (for example, *Abies* and *Picea*) under climate warming[19] were similar in terms of how their photosynthetic responses to warming were shaped by soil moisture availability. When analyses were made for every species independently, the slope of A_{net} to VWC was always steeper in warmed than in ambient plants (Fig. 1 and Extended Data Table 3), and the interaction of warming × VWC was significant ($P < 0.05$ in 10 species, $P = 0.10$ in the other).

Additionally, and as expected because of greater evaporative gradients from warmed plants and soils to the atmosphere[3,8,9,20], the warming treatment reduced soil moisture (Extended Data Fig. 1). Thus, on any given day, warmed plants operated at lower soil moisture levels than ambient plants, moving them to a lower VWC on the A_{net}–VWC relationship than ambient plants. This is illustrated by arrows showing the average VWC of ambient and warmed plants in Fig. 1.

Paralleling the response of A_{net}, leaf diffusive conductance (g_s) decreased in drying soils; it was generally equal or greater in warmed than in ambient plants in moist soils, but similar or lower in warmed than in ambient plants in dry soils (Fig. 2). Moreover, the relationship between g_s and VWC had a steeper slope in the warmed than in the ambient treatment (Fig. 2 and Table 1), the same as for A_{net} (Fig. 1). Evidence suggests that the changes in g_s contributed to the shrinking positive effect of warming on A_{net} as soil water availability decreased (Fig. 1). First, g_s declined proportionally more than A_{net} with increasing soil water deficits (that is, A_{net}/g_s was greater in drier than wetter soils in every species) and the increase in A_{net}/g_s with decreasing soil moisture was larger in warmed compared to ambient plants. Such patterns are consistent with increasing stomatal limitation to A_{net} in drier soils and with greater stomatal limitation in warmed than in ambient plants in drier soils. Second, corroborating this, quantitative estimates of the percentage of limitation of A_{net} by stomatal conductance[25,26] (rather than by biochemical limitations), also increased more steeply with decreasing VWC in warmed than in ambient plots (Extended Data Fig. 4).

A key question is the degree to which the different responses of g_s and A_{net} to VWC for plants in the contrasting warming treatments were influenced by effects of treatments on, or by ambient variation in, other environmental factors such as T_{leaf} and VPG. VWC was very weakly positively correlated with T_{leaf} and unrelated to VPG across all measurement dates (Extended Data Fig. 2); therefore, low soil moisture effects were not confounded by high VPG or high T_{leaf} in this dataset. The differential response of g_s to VWC in warmed versus ambient plants was independent of either VPG or T_{leaf} (no three-way interactions, Table 1). The greater decline of A_{net} with decreasing VWC in warmed than in ambient plants was slightly steeper at higher levels of T_{leaf} and VPG (illustrated by three-way interactions for A_{net} with warming treatment, VWC and either T_{leaf} or VPG, Table 1), but was apparent regardless of VPG or T_{leaf} (Extended Data Fig. 5). Although the relationship of g_s (but not A_{net}) to VPG was nonlinear, replacing VPG with log(VPG) in models in Table 1 only marginally influenced results and did not

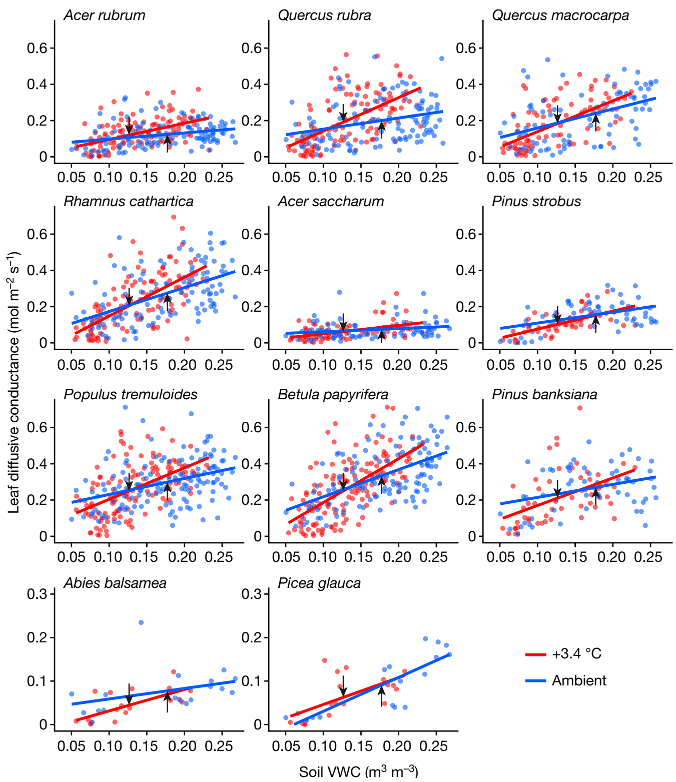


Fig. 2 | Leaf conductance is reduced by drying soils, and more so with simulated climate warming. Leaf diffusive conductance in relation to soil moisture (VWC) by species for ambient (blue) and experimentally warmed (red) plants. Data are from multiple days across three years ($n = 1{,}903$ across species). The slope of g_s versus VWC was significantly steeper in warmed than in ambient plants (Table 1; $F_{1,937} = 6.4$, $P = 0.0113$). The arrows show the median VWC across all measurements for the ambient and warmed plants.

show any interaction of treatment × log(VPG) × VWC, suggesting that nonlinearity of VPG effects did not mask important interactions in the mixed models.

Recent work has shown that under present and projected future climate conditions, canopy surface conductance and evapotranspiration in many biomes, including mesic forests, may be limited by both high vapour pressure deficits (closely related to VPG) and low soil water availability[2]. Our results are consistent with that, as low VWC and high VPG independently constrained A_{net} and g_s (Extended Data Fig. 5).

It is also useful to view these results in the context of the temperature response functions of A_{net}. For both well-hydrated detached leaves[21] and in situ leaves (Extended Data Fig. 2), the broad temperature optima (T_{opt}) of A_{net} for these species was around 22–27 °C. As plants were measured across a wide range of T_{leaf} (95% fell between 13.7 and 36.8 °C, Extended Data Fig. 2), approximately one-third of ambient treatment measurements were made below T_{opt} (for example, $T_{leaf} < 22$ °C) and another third were made above T_{opt} (for example, $T_{leaf} > 29$ °C). Warming by +3.4 °C should have alleviated low temperature limitation for the former and exacerbated high temperature limitations for the latter. The remaining measurements were made when T_{leaf} was near T_{opt} (that is, in the range of 22–29 °C). More influential to the results was that non-optimal VWC induced stomatal closure (Fig. 2), causing a high proportion of leaves to photosynthesize below their capacity at any given T_{leaf} (Extended Data Figs. 2, 4).

Results above clearly demonstrate a more pronounced decline in A_{net} with decreasing VWC in warmed than in ambient plants—congruent with climate-warming stimulation of A_{net} in moist soils and depression of A_{net} in dry soils—and that a more pronounced increase in stomatal limitation of A_{net} of warmed plants played a part. However, this leads to the question of why the shift with declining VWC from biochemically to stomatally limited photosynthesis was steeper in warmed than in ambient plants of all species (Extended Data Fig. 4). We suggest, from several lines of evidence, that a combination of factors drove these responses (Extended Data Fig. 6).

In moist soils, angiosperm species had strong increases in A_{net} and g_s in warmed conditions likely because of both higher carboxylation capacity (greater $V_{cmax\text{-}25}$ in warmed conditions, Extended Data Fig. 3) and higher carbon demand for photosynthate[28], as they grew 23% faster on average in warmed than in ambient conditions[19]. In drier soils, increased stomatal limitation eliminated most of the potential gain that higher $V_{cmax\text{-}25}$ might provide (Extended Data Figs. 3, 6), and perhaps eliminated any warming-induced increase in carbon sink strength. Warmed angiosperm plants also likely had higher dark respiration in the light (as their dark respiration was 20% higher than that of ambient plants[22]) and higher photorespiration[27] at all VWC levels (Extended Data Fig. 6).

The responses of gymnosperms were similar, except that changes in $V_{cmax\text{-}25}$ with warming were less positive even in moist soils; additionally, a negative overall growth response (−26% growth response on average[19]) to warming, coupled with more negative effects of warming on carbon gain when soils were dry, suggests a small warming-induced increase in carbon sink strength at best when soils were wet and a larger decrease when soils were dry (Extended Data Fig. 6). Collectively these factors are likely to have contributed to making the responses of gymnosperms to warming more negative than that of angiosperms at every level of VWC.

Overall, the likely mechanisms suggest that warmed plants did not have greater stomatal sensitivity to soil water deficits as such. Instead, under moist conditions, biochemical limitations to photosynthesis were dominant or co-dominant (Extended Data Fig. 4) and warmed plants had a photosynthetic advantage because of less biochemical limitation (that is, higher realized V_{cmax}), whereas under drier conditions, stomatal limitations became dominant, and any advantage of warming disappeared (and in driest soils, became a hindrance).

The net effect (across the growing season) of warming on photosynthetic carbon gain would be determined by both the shifting effect of warming on A_{net} as it varied with soil water status and the effect of climate warming on soil water status itself. Figure 1 shows the response of warmed versus ambient plants across all levels of soil moisture, that is, comparing the effect of warming on photosynthetic processes at a common soil moisture (and typically not a common date). By contrast, in Fig. 3 we show A_{net} averaged across species in warmed versus ambient plants at a common time, under conditions differing in soil moisture across time and treatments, from dry to wet (representing the 5th, 25th, 50th, 75th and 95th wettest percentiles of VWC among all measurements for each treatment, Fig. 3). Although soils were typically somewhat drier in the warmed treatment, the percentiles (from dry to wet) within each treatment occurred on similar sets of days. Thus, Fig. 3 shows the estimated aggregated effect of both direct physiological effects of warming and indirect soil moisture effects of warming treatments on realized average photosynthetic rates, equally weighted across all 11 species.

The warming treatment had a markedly different effect on A_{net} when soils were dry rather than wet (Fig. 3). For the 11 species, warming under high soil moisture conditions (the 95th percentile of VWC in each treatment) increased A_{net} by 15% on average (Fig. 3). On days with drier conditions, the mean stimulation of A_{net} disappeared; this occurred at around the 65th percentile of VWC on average across the 11 species. Thus, warming increased average A_{net} of the community on only the third of days with highest soil moisture. Species (such as the temperate *Acer* and *Quercus*) with more positive average responses to warming had positive responses for a larger fraction of days and soil water conditions than species with more neutral or negative responses (such as the boreal *Abies*, *Betula*, *Picea* and *Pinus*). On average across species, A_{net} was reduced by the warming treatment by 9%, 18% and 18%, respectively, when soil moisture was at its median, 25th and 5th percentiles. Note that comparisons of A_{net} at the median VWC of

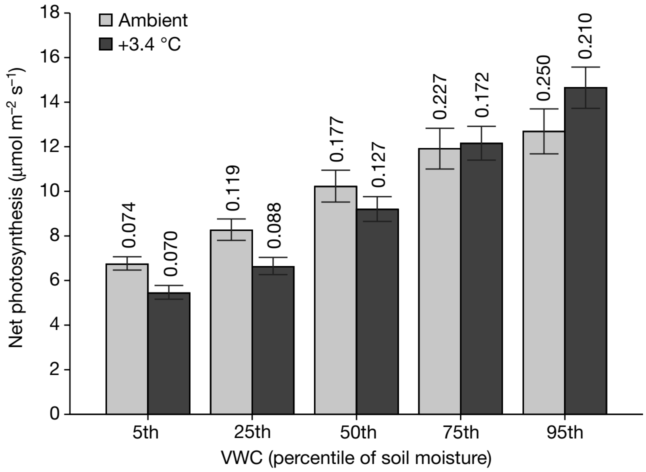


Fig. 3 | Warming stimulates photosynthesis on average in moist soils, but not otherwise. Mean A_{net} (±s.e.m.) of 11 temperate and boreal species in ambient and warmed treatments compared during periods that ranged from dry to wet. Periods represent soil moisture percentiles within treatments across all measurements, from dry to wet (that is, the 5th, 25th, 50th, 75th and 95th wettest percentiles of VWC for each treatment). The percentiles (from dry to wet) occurred on nearly identical days in both treatments. Values represent the predictions for each warming treatment averaged across all 11 species at each VWC level, based on the coefficients for VWC from within-treatment mixed models using VWC, species and their interaction ($n = 996$ for ambient, 995 for warmed; VWC, $P < 0.0001$ in both treatments based on F-tests). The s.e.m. is derived from the standard error of the slope of A_{net} versus VWC within each treatment. Note that the mean VWC by treatment is also shown at each soil moisture percentile above each graph.

ambient and warmed treatments can also be obtained for each species from the arrows in Fig. 1. Results restricted to the nine species measured in two or three years, or to the five species measured in all three years, were generally similar to results for all 11 species: when soil moisture was high, warming increased A_{net}, but whenever substantial soil moisture deficits occurred, warming decreased A_{net} (Extended Data Table 4).

These results provide information on how soil moisture may modulate the effects of climate warming in seasonally cold forest ecosystems, which represent approximately half of global forests[29]. During periods of low soil moisture, stomatal limitation of photosynthesis reduced or eliminated the potential benefit of amelioration of low temperature constraints on photosynthetic kinetics by warming (Figs. 1, 2 and Extended Data Figs. 3, 4, 6). On average, warmed plants had higher g_s and A_{net} than ambient plants when soils were moist (Figs. 1, 2). As soils dried, plants in both treatments showed reduced g_s, but warmed plants of all species had reductions in both g_s and A_{net} that were proportionally higher than in ambient plants. In a warmer future, greater increases in evapotranspiration than in precipitation during the growing season[3] should also reduce soil water stores[9], pushing plants in the future climate further down the 'A_{net}–VWC curve' and further reducing or eliminating positive effects of warming on photosynthetic carbon gain.

Across the three study years, the distribution of soil moisture on the dates of photosynthesis measurements closely matched the distribution of soil moisture across all days (Extended Data Table 2); the three study years were also similar in temperature and precipitation to the 35-year average for these sites (Extended Data Table 1). Thus, the observed responses to experimental warming (Figs. 1–3) are likely to be indicative of responses to future climate warming in northern Minnesota if rainfall patterns are similar to the recent past, and suggest, more generally, that soil water limitations may considerably constrain the realized potential benefits of warming in seasonally cold environments across high latitude forests. Moreover, our results can help to explain observations that climate change to date has had more negative effects on boreal forests in central and western North America than on those further east[5,6,9–11,16,18]. Given higher precipitation and lower evapotranspiration, soils in eastern North American boreal forests are more often moist, and thus higher temperatures are more likely to enhance photosynthesis, whereas in boreal forests in central and western regions, low soil moisture and associated stomatal closure more often constrain photosynthetic carbon gains[3,5,9–11].

Climate warming is likely to extend the season of active photosynthesis, and the effects of increasing CO_2 concentrations on g_s may result in enhanced soil moisture[5,10,15]; both could help to offset the negative effects of soil drying on photosynthesis that result from higher potential evapotranspiration relative to growing season precipitation and from lower soil moisture recharge, resulting from higher rainfall intensity and more run-off[1,3,9–11]. However, the relative magnitude of such offsets is unknown[1,3,9–11]. Furthermore, although the mechanisms that underlie the observations in this experiment should apply to trees of all sizes, larger trees may differ in their sensitivity to drying soils from the juveniles used in this study, influencing the magnitude of soil moisture-related modulation of the effects of climate warming on photosynthesis.

In summary, these results have important implications for the future, arising from two independent but additive mechanisms. First, future warmer conditions will lead to increasingly strong stomatal limitation of photosynthesis in drying soils, such that soil water limitations of historically typical magnitude will eliminate some or all of the increased carbon gain possible from greater photosynthetic capacity. Second, higher evapotranspiration in a warmer world[5,9–11] will result in chronically lower average soil moisture, further reducing net photosynthesis via the same mechanism of decreased stomatal conductance. Thus, low soil moisture will exert a powerful braking effect on, or even reverse, potential benefits of climate warming on tree photosynthesis in mesic, seasonally cold environments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0582-4.

Received: 11 December 2017; Accepted: 16 August 2018;
Published online 3 October 2018.

1. Kao, S. C. & Ganguly, A. R. Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios. *J. Geophys. Res.* **116**, D16119 (2011).
2. Novick, K. A. et al. The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. *Nat. Clim. Change* **6**, 1023–1027 (2016).
3. Seager, R. et al. Dynamical and thermodynamical causes of large-scale changes in the hydrological cycle over North America in response to global warming. *J. Clim.* **27**, 7921–7948 (2014).
4. Moyes, A. B., Castanha, C., Germino, M. J. & Kueppers, L. M. Warming and the dependence of limber pine (*Pinus flexilis*) establishment on summer soil moisture within and above its current elevation range. *Oecologia* **171**, 271–282 (2013).
5. Price, D. T. et al. Anticipating the consequences of climate change for Canada's boreal forest ecosystems. *Environ. Rev.* **21**, 322–365 (2013).
6. Hogg, E. H., Michaelian, M., Hook, T. I. & Undershultz, M. E. Recent climatic drying leads to age-independent growth reductions of white spruce stands in western Canada. *Glob. Change Biol.* **23**, 5297–5308 (2017).
7. IPCC. *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).
8. Sherwood, S. & Fu, Q. A drier future? *Science* **343**, 737–739 (2014).
9. Wang, Y., Hogg, E. H., Price, D. T., Edward, J. & Williamson, T. Past and projected future changes in moisture conditions in the Canadian boreal forest. *Forest. Chron.* **90**, 678–691 (2014).
10. Girardin, M. P. et al. No growth stimulation of Canada's boreal forest under half-century of combined warming and CO_2 fertilization. *Proc. Natl Acad. Sci. USA* **113**, E8406–E8414 (2016).
11. D'Orangeville, L. et al. Northeastern North America as a potential refugium for boreal forests in a warming climate. *Science* **352**, 1452–1455 (2016).
12. Wong, S. C., Cowan, I. R. & Farquhar, G. D. Stomatal conductance correlates with photosynthetic capacity. *Nature* **282**, 424–426 (1979).
13. Bloom, A. J., Chapin, I. F. S. & Mooney, H. A. Resource limitation in plants—an economic analogy. *Annu. Rev. Ecol. Syst.* **16**, 363–392 (1985).
14. Rastetter, E. B. & Shaver, G. R. A model of multiple element limitation for acclimating vegetation. *Ecology* **73**, 1157–1174 (1992).

15. Buermann, W., Bikash, P. R., Jung, M., Burn, D. H. & Reichstein, M. Earlier springs decrease peak summer productivity in North American boreal forests. *Environ. Res. Lett.* **8**, 024027 (2013).
16. Ma, Z. et al. Regional drought-induced reduction in the biomass carbon sink of Canada's boreal forests. *Proc. Natl Acad. Sci. USA* **109**, 2423–2427 (2012).
17. Moyes, A. B., Germino, M. J. & Kueppers, L. M. Moisture rivals temperature in limiting photosynthesis by trees establishing beyond their cold-edge range limit under ambient and warmed conditions. *New Phytol.* **207**, 1005–1014 (2015).
18. Peng, C. et al. A drought-induced pervasive increase in tree mortality across Canada's boreal forests. *Nat. Clim. Change* **1**, 467–471 (2011).
19. Reich, P. B. et al. Geographic range predicts photosynthetic and growth response to warming in co-occurring tree species. *Nat. Clim. Change* **5**, 148–152 (2015).
20. Rich, R. L. et al. Design and performance of combined infrared canopy and belowground warming in the B4WarmED (Boreal Forest Warming at an Ecotone in Danger) experiment. *Glob. Change Biol*. **21**, 2334–2348 (2015).
21. Sendall, K. M. et al. Effects of experimental forest warming on photosynthetic temperature optima of temperate and boreal tree species. *Glob. Change Biol*. **21**, 1342–1357 (2015).
22. Reich, P. B. et al. Boreal and temperate trees show strong acclimation of respiration to warming. *Nature* **531**, 633–636 (2016).
23. Rawls, W. J., Brakensiek, D. L. & Sexton, K. E. Estimation of soil water properties. *Trans. ASAE* **25**, 1316–1320 (1982).
24. Campbell, G. S. & Norman, J. M. *An Introduction to Environmental Biophysics* (Springer-Verlag, New York, 1998).
25. Sharkey, T. D. Photosynthesis in intact leaves of C_3 plants: physics, physiology and rate limitations. *Bot. Rev*. **51**, 53–105 (1985).
26. Jones, H. G. Partitioning stomatal and non-stomatal limitations to photosynthesis. *Plant Cell Environ*. **8**, 95–104 (1985).
27. Sage, R. F. & Kubien, D. S. The temperature response of C_3 and C_4 photosynthesis. *Plant Cell Environ*. **30**, 1086–1106 (2007).
28. Körner, C. Paradigm shift in plant growth control. *Curr. Opin. Plant Biol*. **25**, 107–114 (2015).
29. FOA. *State of the World's Forests* http://www.fao.org/docrep/013/i2000e/i2000e00.htm (FOA, 2011).

Acknowledgements This research was supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research award DE-FG02-07ER64456; Minnesota Agricultural Experiment Station MIN-42-030 and MIN-42-060; the Minnesota Department of Natural Resources; and the College of Food, Agricultural, and Natural Resources Sciences and Wilderness Research Foundation, University of Minnesota. Assistance with experimental operation and data collection was provided by K. Rice, C. Buschena, C. Zhao, H. Jihua and numerous summer interns. We thank D. Ellsworth, Ch. Messier, J. Drake and B. Medlyn for helpful comments on the manuscript.

Reviewer information *Nature* thanks M. Mencuccini, M. Ryan and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.B.R., R.A.M., R.L.R. and S.E.H. designed the overall experiment. P.B.R., R.A.M. and A.S. designed the specific study reported herein. R.L.R. designed the warming system, R.L.R. and A.S. implemented the warming system, and A.S. and K.M.S. coordinated the day-to-day field measurements. P.B.R. coordinated the overall experiment and this specific study and analysed the data. P.B.R. wrote the first draft and jointly wrote subsequent drafts of the manuscript with the other co-authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0582-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0582-4.
Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.B.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

The experiment is located at two University of Minnesota field stations; the Cloquet Forestry Center, Cloquet MN (46° 40′ 46″ N, 92° 31′ 12″ W, 382 m a.s.l., 4.8 °C mean annual temperature, 783 mm mean annual precipitation) and the Hubachek Wilderness Research Center, Ely, MN (47° 56′ 46″ N, 91° 45′ 29″ W, 415 m a.s.l., 2.6 °C mean annual temperature, 726 mm mean annual precipitation)[19,20]. At both sites, treatments were positioned in relatively open (recently cleared) overstory conditions. The overall experimental design was a 2 (site) × 2 (treatment) factorial experiment, with six replicates of each for a total of 24 circular 3-m diameter plots; with seedlings of 11 focal species planted in every plot. Treatments included two levels of simultaneous open-air plant and soil warming (ambient, +3.4 °C); warming was accomplished with infrared lamp heaters and soil heating cables (dummy lamps and cables in the ambient plots). Warming was implemented from early spring to late fall each year in open-air plots (that is, without chambers) via a feedback control that acts concurrently and independently at the plot scale to maintain a fixed temperature differential from ambient conditions above- and belowground. On average, we achieved 24-h per day average warming of +3.4 °C (during April–November) and midsummer midday (09:00–15:00 during June–September) aboveground warming of +2.9 °C across the 2009–2011 growing seasons[19,20]. Plant and soil temperature and soil moisture (0–20 cm depth) were measured continuously and recorded hourly in every plot throughout the study. Plant surface temperature was measured with infrared thermometers mounted above the plant canopy in every plot (IRR-P: Apogee Instruments Inc.). Volumetric water content from 0 to 20 cm depth was measured in each plot using a 30-cm Campbell Scientific CS-616 probe inserted at 45°. VWC (m^3 H_2O per m^3 soil) was monitored hourly in all plots and corrected[20] for soil textural and temperature differences using a Campbell Scientific method for user-specific calibration of water reflectometers (Model CS616). Both sites have well-drained, coarse-textured upland soils[19,20]. In mid-continental boreal and temperate biomes, climate change will increase plant and air temperatures, and the associated increases in VPG and evapotranspiration are likely to more than offset any increase in total atmospheric water vapour or precipitation, resulting in increased soil water deficits[3,7–10].

In 2008, 11 juveniles of each of 11 tree species were planted into existing low shrub, herb and fern vegetation in every plot (around 2,900 juveniles; average of approximately 3-year-old plants in 2009). The 11 species include six native broadleaf (*Acer rubrum*, *Acer saccharum*, *Betula papyrifera*, *Populus tremuloides*, *Quercus macrocarpa* and *Quercus rubra*), one naturalized broadleaf (*Rhamnus cathartica*) and four native needle leaved (*Abies balsamea*, *Picea glauca*, *Pinus banksiana* and *Pinus strobus*) species, all of which are present in the ecotonal region. Local ecotypes (collected between 46° 0′ and 48° 30′ N latitude in northeastern Minnesota) of all species except *Rhamnus* were planted from material obtained from two Minnesota Department of Natural Resources nurseries in northern Minnesota. *Rhamnus* seedlings were transplants dug up from forests in north-central Minnesota.

In situ measurements of light-saturated net photosynthesis (A_{net}) and leaf diffusive conductance (g_s) were made using six Li-Cor 6400 portable photosynthesis systems (Li-Cor). Simultaneous leaf temperature measurements were made for most species using the internal fine wire thermocouple located in the bottom of the 2 × 3-cm^2 Li-Cor leaf chamber (6400-02B LED) and directly touching the leaf during the measurement. However, for two conifers (balsam fir and spruce), we used a conifer chamber LED light source (6400-22L) and leaf temperature was calculated based on energy balance (for details see Li-Cor 6400XT manual; Li-Cor). Leaf temperatures measured in the cuvette and canopy surface temperatures (measured independently with infrared thermometers, as described above) were strongly correlated. Cuvette leaf temperatures were usually around 2 °C higher than canopy temperature. This is largely because the cuvette and the enclosed leaf warmed up from being in the sun; additionally, leaves were selected for photosynthesis from upper canopy leaves in sunlit positions, whereas part of the surface of the plant canopy sensed by the infrared thermometers was often in partial shade. Measurements were made throughout the growing seasons (June–September) of 2009–2011. A total of 2,052 measurements of A_{net} and 1,964 of g_s were made on a total of 1,338 individuals on 54 dates across species, treatments, sites and time (1,991 and 1,903 measurements, respectively were made with matching soil VWC data). Individuals were three- to five-years old at the time of measurements. Measurements were made in morning or early afternoon (that is, typically between 08:30–14:00 solar time). Not all species were measured each year owing to the time-consuming nature of the measurements (five species were measured in all three years, four in two years and two in one year). On every measurement date, any species included in that sampling was measured equally across contrasting warming treatments. Individuals to be sampled were chosen randomly from those not previously sampled. Every measurement was made on a unique leaf. Over the three years, individual plants were usually measured once ($n = 839$) or twice ($n = 338$), but owing to low survival in some species, other individuals were measured three ($n = 121$), four ($n = 30$), five ($n = 6$) or six ($n = 4$) times. Fully expanded, healthy upper canopy leaves were sampled from individuals in both ambient and +3.4 °C treatments at both sites. Light was maintained in the leaf chamber at saturating levels using the LED light source. Airflow was set at 500 μmol s^{-1} and CO_2 reference concentrations were set at 400 μmol mol^{-1}.

Estimates of $V_{cmax-25}$ from the one-point method[30] and estimates of the percentage of stomatal limitation[25,26] of A_{net} were also made. For data from other years for which full A–C_i curves were measured, calculated $V_{cmax-25}$ from the one-point method from single points of those A–C_i curves very closely matched (near 1:1 line, $R^2 = 0.96$) the $V_{cmax-25}$ values estimated from the entire curves, strongly supporting the appropriateness of the one-point method for our field measurements for this set of species. The percentage of stomatal limitation was taken as the percentage reduction in A_{net} from the maximal rate estimated with no stomatal limitation (A_{gmax}). A_{gmax} was estimated (for each species in both treatments) in three ways: (1) based on calculations from A–C_i curves of nine of the eleven species made in later years of the study on a separate cohort of plants; (2) based on the 95th percentile of A_{net} measurements from the current study, and (3) based on the A_{gmax} estimates from the A–C_i curves, adjusted to reflect realized A_{net} in the current study using the correlation of values from 1 and 2. For method 1, we used the relationship between A–C_i curves and the field 95th percentile A_{net} for nine species to estimate A_{gmax} for the two species without A–C_i curves. The overall patterns shown in each panel of Extended Data Fig. 4 are nearly identical using any of the three metrics. We used metric 3, because it combined independent estimates of net photosynthetic rates from outside of this study, with maximal rates that better reflected realized rates in the study (and thus resulted in fewer values below zero for the percentage of stomatal limitation). We recognize the impossibility of negative values for the percentage of stomatal limitation, but retained them for statistical purposes.

A mixed model was used to compare A_{net} and g_s to treatment combinations, soil moisture conditions, VPG and leaf temperature. Models included the following independent variables: species, warming treatment, VWC (on the day the gas exchange measurement was made), VPG, T_{leaf} and all interactions (up to four-way) among variables. Plot, block and site were added to each model as a random effect. Models were also run separately for the subset of nine species measured in at least two years (Extended Data Table 4), for the five species measured in all three years (Extended Data Table 4) and for each species individually (Extended Data Table 3). Results were similar across these different models. Moreover, comparisons across species on common dates were made in three different ways. First, we used coefficients from mixed models for each temperature treatment to estimate A_{net} across a range of VWC percentiles (Fig. 3). Second, we ran mixed models, including species, treatments and VWC bin classes to develop LSMEANs for all species × treatment × VWC bin combinations. Third, we averaged raw species means for VWC bin classes across treatments. All three approaches resulted in similar outputs.

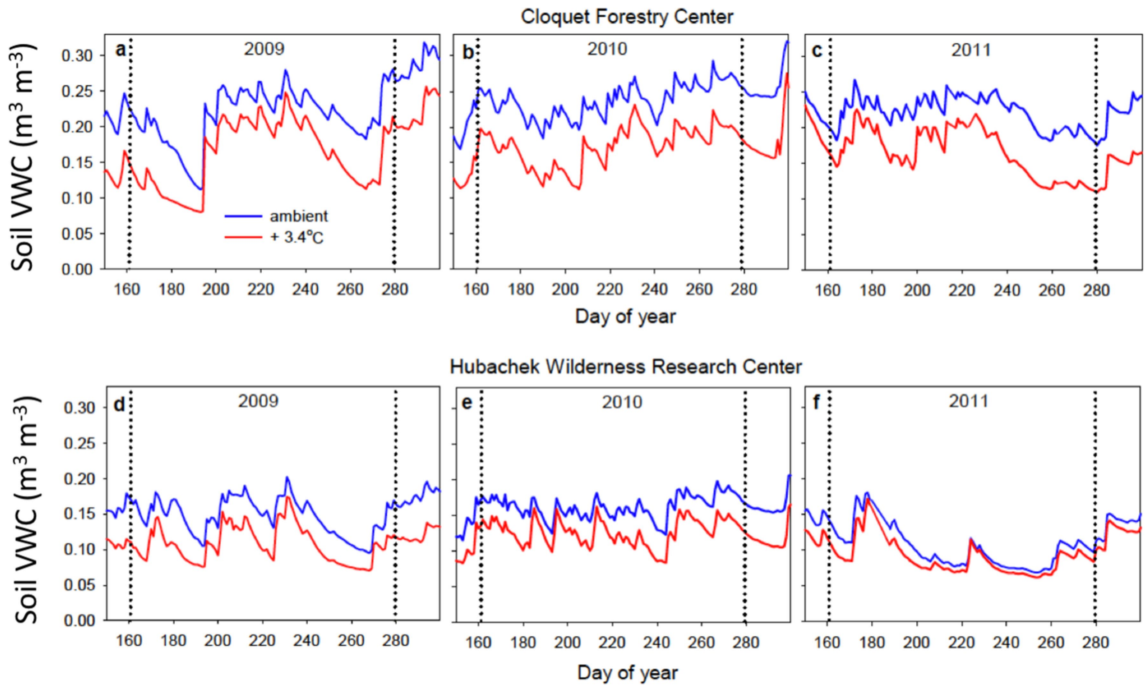
The three experimental years were typical of long-term climate (Extended Data Table 1); moreover, over the three years, the dates when leaf physiological measurements were made were well-distributed from early June to late September (between day of year 162 and 269), and represented a similar range of frost-free temperatures and soil moisture as occurred across that growing season period in 2009–2011 (Extended Data Table 2).There was no evidence that mid-summer, which is warmer, was on average drier during these three particular years, nor did periods of low VWC occur in times of high VPG. As a result, there was no confounding of soil moisture deficits with leaf or air temperatures or VPG during our study; thus, physiological effects related to low soil moisture should have been largely independent of effects of air temperature (or VPG).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

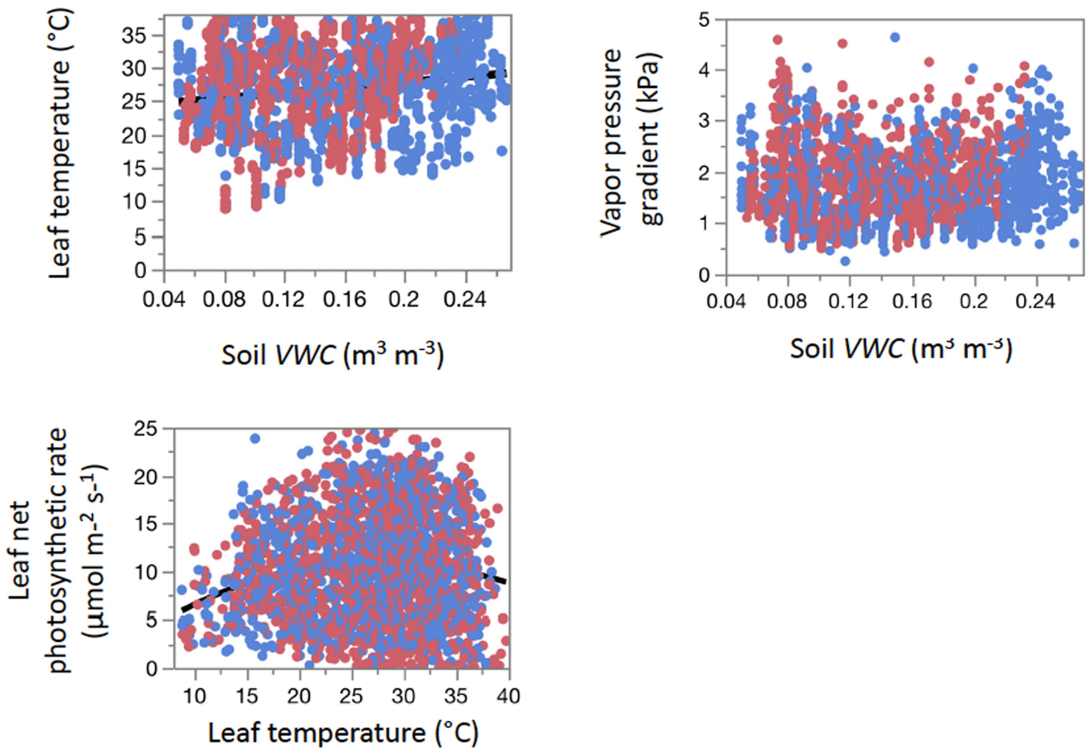
Data availability

The data reported in this paper are available from the Environmental Data Initiative (EDI) at https://doi.org/10.6073/pasta/258239f68244c959de0f97c922ac313f.

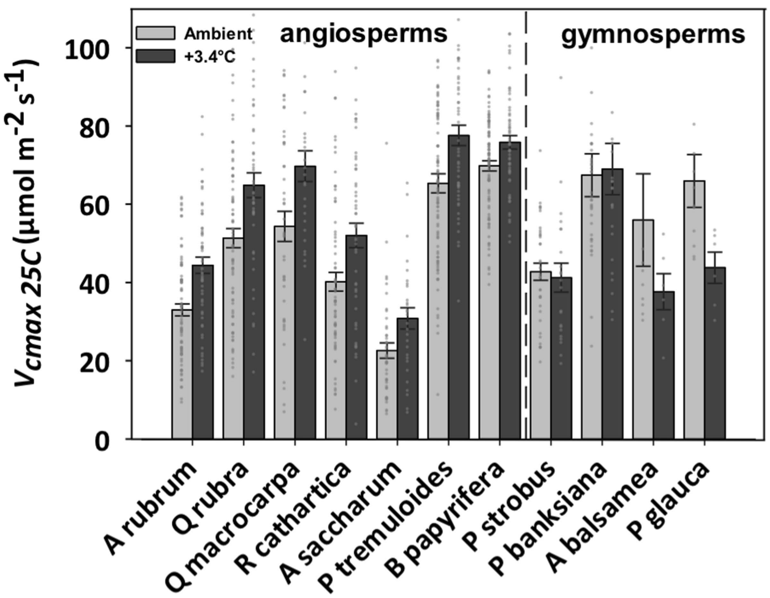
30. De Kauwe, M. G. et al. A test of the 'one-point method' for estimating maximum carboxylation capacity from field-measured, light-saturated photosynthesis. *New Phytol.* **210**, 1130–1144 (2016).



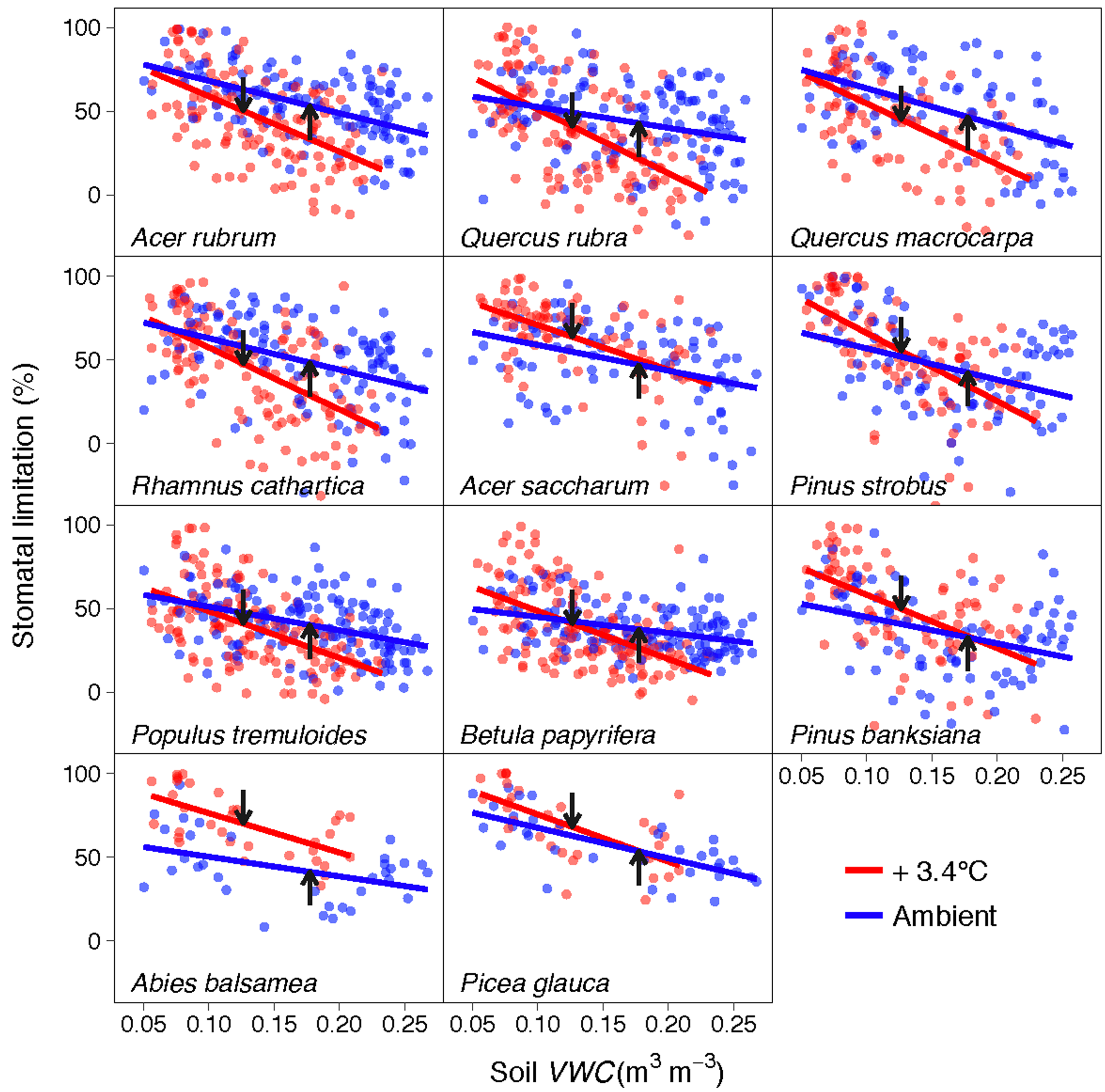
Extended Data Fig. 1 | Soil water (VWC) in relation to day of year. **a–f**, VWC (m^3 m^{-3}; 0–20 cm depth) was averaged by day, variation shown daily across the season among treatments, sites and years. Daily values represent means among all plots within a treatment at each site. Measurements were logged continuously, recorded hourly, thus a total of approximately 3,600 measurements for each of the 24 plots in each year for the time period are shown. Vertical dashed lines show the range of dates during which photosynthetic measurements were made.



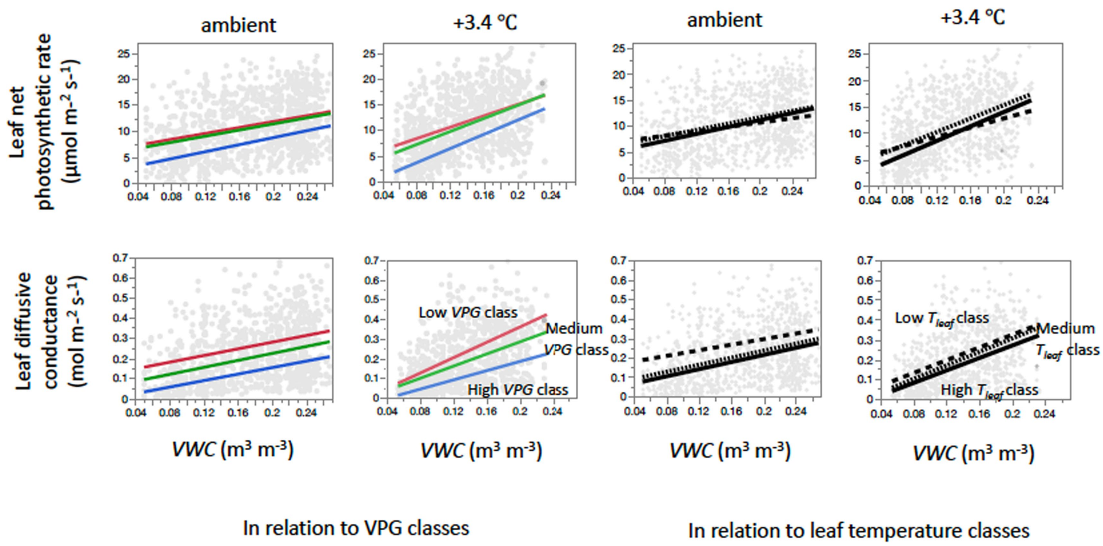
Extended Data Fig. 2 | Range of temperature, evaporative demand (VPG) and soil moisture across the three growing seasons during gas exchange measurements. Top, average leaf temperature and VPG for all gas exchange measurements across the three years in relation to soil water (VWC). There was no significant correlation between VPG and VWC over the three-year period ($P > 0.30$); there was a significant correlation ($R^2 = 0.03$, $P < 0.001$) between leaf temperature and VWC across warming treatments. Bottom, net photosynthetic rate in relation to leaf temperature (polynomial fit all data pooled, $R^2 = 0.02$, $P < 0.001$). Blue, ambient; red, +3.4 °C. Sample sizes, approximately 1,989–2,050, around half in each warming treatment. A few data points are out of the *y*-axis range and therefore not visible.



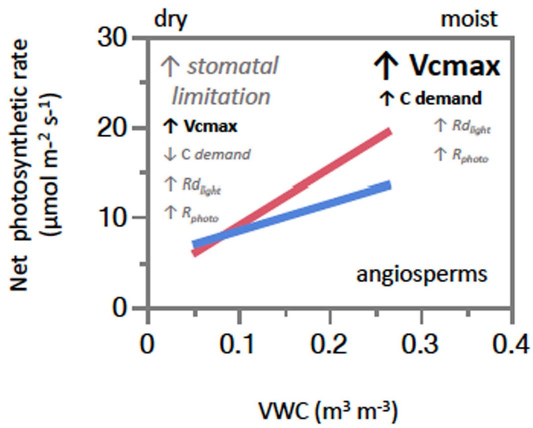
Extended Data Fig. 3 | Maximum biochemical photosynthetic capacity in moist soils. Mean (±s.e.m.) maximum carboxylation capacity ($V_{\text{cmax-25}}$, μmol m^{-2} s^{-1}) at 25 °C of 11 gymnosperm and angiosperm trees species in ambient (grey) and +3.4 °C experimentally warmed (black) treatments for days with moist soils (data are shown for the highest half of VWC observations, those with VWC > 0.148). Species within groups are arranged from left to right from most temperate to most boreal distribution (as in Fig. 1). Data are from multiple days across three years and otherwise averaged across the spectrum of moist soil water availability. Individual measurements are shown as small grey dots. Sample sizes by species for ambient, +3.4 °C: *A. rubrum*, 78, 55; *Q. rubra*, 75, 47; *Q macrocarpa*, 43, 28; *R. cathartica*, 69, 48; *A. saccharum*, 44, 29; *P. tremuloides*, 92, 50; *B. papyrifera*, 91, 56; *P. strobus*, 36, 22; *P banksiana*, 36, 24; *A. balsamea*, 10, 6; *P glauca*, 11, 6. A few data points are out of the *y*-axis range and are therefore not visible.

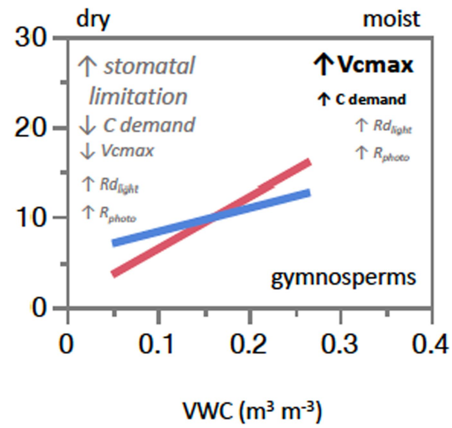


Extended Data Fig. 4 | Percentage of stomatal limitation of net photosynthesis in relation to soil moisture (VWC) by species for ambient and experimentally warmed plants. The percentage of stomatal limitation was calculated according to previous studies[25,26]. Data are from multiple days across three years ($n = 1{,}991$ across species). In a full model analogous to those used in Table 1, the slope of the percentage of stomatal limitation versus VWC was significantly steeper in warmed (red) than ambient (blue) plants (interaction of VWC × warming treatment, $F_{1,593} = 38.1$, $P < 0.0001$). The arrows show the median VWC across all measurements for the ambient and warmed plants of each species. Species are arranged from top to bottom by their geographical ranges (temperate species in top two rows, boreal in bottom two rows).



Extended Data Fig. 5 | Relationships of net photosynthesis and leaf conductance to soil water content for different VPG classes and leaf temperatures. Relationships are shown for two temperature treatments, for three VPG classes (left four panels) and three leaf temperature classes (right four panels). Data are pooled across all species and show the regression line for A_{net} and g_s in relation to VWC in three VPG classes (0.4–1.6, 1.6–2.8, 2.8–4.0 kPa; red, green and blue lines, respectively) and for ambient and warmed (+3.4 °C) treatment plants; and in relation to VWC in three T_{leaf} classes (8–20, 20–32, 32–38 °C; dashed, dotted, and solid black lines, respectively) for ambient and warmed (+3.4 °C) treatment plants. Sample sizes in each panel, around 950–995. A few data points are out of the *y*-axis range and therefore not visible.





Extended Data Fig. 6 | Conceptual illustration of mechanisms that influence the effect of climate warming on the response of realized A_{net} and soil water content (VWC). Schematics are shown for angiosperms (left) and gymnosperms (right). Red lines indicate warmed treatment plants, blue lines ambient plants. The regression lines are pooled for all seven angiosperms and all four gymnosperms, at each warming treatment. The arrows show the direction of the effect of warming treatment on specific factors and the size of the letters indicates the relative magnitude of those effects on A_{net}. Bold fonts in black indicate changes that increase A_{net} in warmed plants relative to ambient plants, italic fonts in grey indicate changes that decrease A_{net} in warmed plants relative to ambient plants. For angiosperms in moist soils, warmed plants exhibit large increases in V_{cmax} and in carbon demand (from 23% higher growth[19]) that far outweigh the likely modest increases in dark respiration in the light (R_{light})[22] and in photorespiration (R_{photo})[27], to result in large increases in A_{net}. For angiosperms in dry soils, however, experimental warming results in lower water availability that slows growth, reducing carbon demand in warmed (compared to ambient) plants. In dry soils, warming also increases stomatal limitation of photosynthesis (perhaps due in part to slightly higher VPG in warmed plots), and constrains the magnitude of positive effects of V_{cmax} on A_{net}. The combination of increased R_{light} and R_{photo} and reduced carbon demand, slightly outweigh increased V_{cmax} and result in slightly reduced A_{net} in warmed compared to ambient angiosperms. The responses of gymnosperms are similar, except that changes in V_{cmax} with warming are less positive (than in angiosperms) in moist soils and negative in dry soils; additionally, the negative overall growth response (−26% growth response on average[19]) to warming suggests at most a small warming-induced increase in carbon sink strength when soils are wet and a larger decline when soils are dry. Collectively these factors make the A_{net} response of gymnosperms to warming more negative than that of angiosperms at every VWC level. Additionally (not shown in this conceptual figure, see Fig. 1), climate warming leads to higher evapotranspiration and thus more pronounced soil drying, therefore warmed plants operate at lower levels of VWC on average (Fig. 1) and at the vast majority of points in time (Extended Data Fig. 1), promoting the tendency of warmed plants to have lower A_{net} on average than ambient plants (Fig. 3).

Extended Data Table 1 | Annual climate means for the two sites before and during the experiment

	Prior to experiment (1973-2008)		During experiment (2009-2011)	
Weather Station	Mean annual precipitation (mm) (SD)	Mean annual temperature (°C) (SD)	Mean annual precipitation (mm) (SD)	Mean annual temperature (°C) (SD)
Cloquet	783.4 (138.5)	4.8 (1.0)	776.1 (117.8)	5.1 (0.8)
Tower (Ely)	725.9 (135.5)	2.6 (1.0)	615.7 (123.6)	3.6 (0.5)

The Tower (Ely) weather station is 43 km from the research site. The Cloquet weather station is 3 km from the research site. Data are mean ± s.d. among years.

Extended Data Table 2 | VWC percentiles for measurement dates and all dates

Dataset	Treatment	5%	25%	mean	50%	75%	95%
Measurement days	ambient	0.080	0.131	0.172	0.181	0.227	0.247
All days	ambient	0.082	0.146	0.182	0.182	0.229	0.259
Measurement days	+3.4 ºC	0.071	0.089	0.132	0.124	0.171	0.203
All days	+3.4 ºC	0.072	0.103	0.137	0.132	0.173	0.211

VWC (0–20 cm depth) values are recorded hourly in every plot across the shown time periods, averaged by day by treatment and then assessed by percentiles for measurement days and all days in both treatments. VWC percentiles are shown for days when leaf gas exchange measurements were made (measurement days) versus all days between (and including) day of year 162 (11 June) and 269 (26 September) from 2009 to 2011 across sites (all days). Results show that for both warming treatments, soil moisture conditions across the measurement days were well-matched to the average conditions across the three growing seasons.

Extended Data Table 3 | Species-specific models of photosynthesis in relation to warming treatment and soil moisture

Species	n	*VWC*	Warm	Warm x *VWC*	R^2	*a* amb/ +3.4	*b* amb/ +3.4
Abies balsamea	62	**<0.0001**	**0.0241**	**0.0227**	0.42	4.8/0.1	12.6/35.2
Acer rubrum	245	**<0.0001**	**<0.0001**	**0.0080**	0.30	2.0/1.4	26.9/47.3
Acer saccharum	166	**<0.0001**	**0.0123**	**0.0031**	0.23	2.3/0.3	12.6/32.7
Betula papyrifera	271	**<0.0001**	0.1199	**<0.0001**	0.31	11.5/4.9	24.7/72.4
Picea glauca	59	**<0.0001**	0.5125	0.1008	0.62	2.0/-1.2	38.7/56.9
Pinus banksiana	136	**<0.0001**	0.9653	**0.0108**	0.25	9.6/2.8	25.8/68.9
Pinus strobus	132	**<0.0001**	0.4752	**0.0326**	0.38	2.9/0.0	32.2/53.8
Populus tremuloides	271	**<0.0001**	**0.0001**	**0.0047**	0.24	8.4/5.9	34.2/64.9
Quercus macrocarpa	175	**<0.0001**	**0.0001**	**0.0394**	0.33	3.0/2.1	46.5/73.8
Quercus rubra	242	**<0.0001**	**<0.0001**	**<0.0001**	0.27	6.5/1.9	22.6/73.1
Rhamnus cathartica	231	**<0.0001**	**<0.0001**	**0.0014**	0.29	3.5/1.1	32.5/67.6

Species-specific tests (*P* values) of A_{net} versus VWC, warming treatment and their interaction are shown, including the sample size (*n*) and R^2 of the full model. Intercepts (*a*) and slopes (*b*) for A_{net} versus VWC in ambient (amb) and warmed (+3.4 °C; warm) treatments were examined separately. Relationships are shown in Fig. 1. For all species, the relationship of A_{net} to VWC was positive. For species for which the relationship of A_{net} to warming was significant, it was positive, except it was negative for *A. balsamea*. All full models were significant ($P < 0.0001$). Sample sizes for the percentage of stomatal limitation (Extended Data Fig. 4) were nearly identical to those shown here for A_{net} versus VWC (within 1% for each species). Sample sizes for $g_s \times$ VWC shown in Fig. 2 are identical, except for *A. balmasea* ($n = 38$), *P. glauca* ($n = 36$), *P. banksiana* ($n = 117$), *P. strobus* ($n = 112$) and *R. cathartica* ($n = 228$). For Figs. 1, 2 and Extended Data Fig. 4 roughly half of the measurements were in each warming treatment.

Extended Data Table 4 | Summaries of mixed model analyses for species examined across two or three years

A. Nine species measured in two or three years

Source of variance	(A_{net})		(g_s)	
	F	P>F	F	P>F
Species	158.03	**<0.0001**	103.17	**<0.0001**
Warm	47.65	**<0.0001**	11.26	**0.0035**
Species*Warm	4.65	**<0.0001**	1.40	0.1923
Soil water	492.98	**<0.0001**	527.41	**<0.0001**
Soil water *Species	3.64	**0.0003**	14.24	**<0.0001**
Soil water *Warm	59.18	**<0.0001**	18.90	**<0.0001**
Soil water*Species*Warm	1.23	0.2755	1.11	0.3504
Full model R^2	0.56		0.52	

B. Five species measured in three years

Source of variance	(A_{net})		(g_s)	
	F	P>F	F	P>F
Species	177.60	**<0.0001**	100.91	**<0.0001**
Warm	55.31	**<0.0001**	19.46	**0.0004**
Species*Warm	3.72	**0.0051**	0.92	0.4529
Soil water	341.67	**<0.0001**	463.80	**<0.0001**
Soil water *Species	1.20	0.3098	14.63	**<0.0001**
Soil water *Warm	53.15	**<0.0001**	22.14	**<0.0001**
Soil water*Species*Warm	1.62	0.1675	0.99	0.4106
Full model R^2	0.53		0.49	

a, **b**, Analyses for A_{net} and g_s in relation to +3.4 °C warming treatment (Warm), species, soil water (VWC) and their interactions for nine species measured in at least two years (**a**) and five species measured in all three years (**b**). Plot, block and site were included as random effects in the models. Data for **a** are a subset of nine species (all species except *A. balsamea* and *P. glauca*) measured in at least two years ($n = 1{,}870$ for A_{net}; 1,829 for g_s). Data for **b** are a subset of five species (*A. rubrum*, *B. papyrifera*, *P. tremuloides*, *Q. rubra* and *R. cathartica*) measured in all three years ($n = 1{,}260$ for A_{net}; 1,259 for g_s). Both models were significant at $P < 0.0001$.

https://doi.org/10.1038/s41586-018-0566-4

Common genetic variants contribute to risk of rare severe neurodevelopmental disorders

Mari E. K. Niemi[1], Hilary C. Martin[1], Daniel L. Rice[1], Giuseppe Gallone[1], Scott Gordon[2], Martin Kelemen[1], Kerrie McAloney[2], Jeremy McRae[1], Elizabeth J. Radford[1,3], Sui Yu[4], Jozef Gecz[5,6], Nicholas G. Martin[2], Caroline F. Wright[7], David R. Fitzpatrick[8], Helen V. Firth[1,9], Matthew E. Hurles[1] & Jeffrey C. Barrett[1]*

There are thousands of rare human disorders that are caused by single deleterious, protein-coding genetic variants[1]. However, patients with the same genetic defect can have different clinical presentations[2–4], and some individuals who carry known disease-causing variants can appear unaffected[5]. Here, to understand what explains these differences, we study a cohort of 6,987 children assessed by clinical geneticists to have severe neurodevelopmental disorders such as global developmental delay and autism, often in combination with abnormalities of other organ systems. Although the genetic causes of these neurodevelopmental disorders are expected to be almost entirely monogenic, we show that 7.7% of variance in risk is attributable to inherited common genetic variation. We replicated this genome-wide common variant burden by showing, in an independent sample of 728 trios (comprising a child plus both parents) from the same cohort, that this burden is over-transmitted from parents to children with neurodevelopmental disorders. Our common-variant signal is significantly positively correlated with genetic predisposition to lower educational attainment, decreased intelligence and risk of schizophrenia. We found that common-variant risk was not significantly different between individuals with and without a known protein-coding diagnostic variant, which suggests that common-variant risk affects patients both with and without a monogenic diagnosis. In addition, previously published common-variant scores for autism, height, birth weight and intracranial volume were all correlated with these traits within our cohort, which suggests that phenotypic expression in individuals with monogenic disorders is affected by the same variants as in the general population. Our results demonstrate that common genetic variation affects both overall risk and clinical presentation in neurodevelopmental disorders that are typically considered to be monogenic.

We carried out a genome-wide association study (GWAS) in 6,987 patients with severe neurodevelopmental disorders and 9,270 ancestry-matched controls, using common variants with a minor allele frequency $\geq 5\%$ (Fig. 1, Extended Data Fig. 1, Supplementary Tables 1, 2 and Methods). The patients were recruited by senior clinical geneticists in the UK and Ireland as part of the Deciphering Developmental Disorders (DDD) study[6,7]. They all had at least one abnormality that affects the morphology or physiology of the central nervous system, and to be recruited to the study their clinical features were sufficiently severe that their disorder was thought likely to be monogenic. In addition to neurodevelopmental defects—for example, global developmental delay, intellectual disability, cognitive impairment or learning disabilities in 86% of the cohort, and autism spectrum disorders in 16% of the cohort (Fig. 2a)—88% of the recruited patients also had abnormalities in at least one other organ system (Fig. 2b and Extended Data Table 1).

We did not find any single-variant associations at genome-wide significance (Extended Data Fig. 2a), which was unsurprising given the heterogeneity of our clinical phenotype and the presumption that these disorders are monogenic. We did, however, observe a modest inflation in the test statistics ($\lambda = 1.097$, Extended Data Fig. 2b), which could indicate either residual bias between cases and controls or a polygenic contribution of common variants to disease risk. We therefore estimated common-variant heritability using linkage-disequilibrium score (LD score) regression[8], which can differentiate between these two possibilities, and found that 7.7% (standard error (s.e.) = 2.1%) of variance in risk (on the liability scale) for neurodevelopmental disorders in our sample was attributable to common genetic variants, when assuming a population prevalence of 1% (Methods). This common variant heritability estimate (h^2) is similar to that which has been reported for common disorders such as autism ($h^2 = 11.8\%$, s.e. = 1.0%)[9] and major depressive disorder ($h^2 = 8.9\%$, s.e. = 0.4%)[10]. To replicate this signal, we analysed an independent set of 728 trios recruited as part of the same study, but who were not in the initial GWAS. We calculated polygenic scores for each individual by summing the genetic effects across all independent variants from our discovery GWAS (Fig. 1 and Methods). We then performed a polygenic transmission disequilibrium test[11], which compares the mean parental polygenic scores to those of the affected children. We found that our neurodevelopmental disorder risk score was over-transmitted in these trios ($P = 0.0035$, $t = 2.48$, degrees of freedom = 727, one-sided t-test), which confirms that common variants contribute to risk of disorders widely presumed to be monogenic.

Previous studies have shown that the risk of more common neuropsychiatric disorders—for example, schizophrenia and bipolar disorder[12,13]—and variation in other brain-related traits, including educational attainment[13], is driven in part by shared common genetic effects. We therefore used the LD score method[14] to test for genetic correlation between our GWAS of neurodevelopmental disorders and available GWAS data for common neuropsychiatric disorders, cognitive and educational traits and anthropometric traits, as well as negative-control diseases that have well-powered GWAS but are not related to neurodevelopment. We found that genetic risk for neurodevelopmental disorders was significantly negatively correlated with genetic predisposition (as measured by Spearman's g) to higher educational attainment[15] ($r_g = -0.49$, s.e. = 0.08, $P = 5.3 \times 10^{-10}$) and intelligence[16] ($r_g = -0.44$, s.e. = 0.10, $P = 2.2 \times 10^{-5}$), and positively correlated with genetic risk of schizophrenia ($r_g = 0.28$, s.e. = 0.07, $P = 2.7 \times 10^{-5}$) (Fig. 3 and Extended Data Table 2). None of the anthropometric or negative-control traits were significantly genetically correlated with our data, after accounting for multiple testing. We also used partitioned LD score regression[17] to show that heritability of neurodevelopmental disorders was nominally significantly enriched

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [2]QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [3]Department of Paediatrics, University of Cambridge, Cambridge, UK. [4]Department of Genetics and Molecular Pathology, SA Pathology, Women's and Children's Hospital, Adelaide, South Australia, Australia. [5]Adelaide Medical School and Robinson Research Institute, Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, South Australia, Australia. [6]South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia. [7]University of Exeter Medical School, Institute of Biomedical and Clinical Science, RILD, Royal Devon & Exeter Hospital, Exeter, UK. [8]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh, UK. [9]Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. *e-mail: jeff.barrett@genomicsplc.com

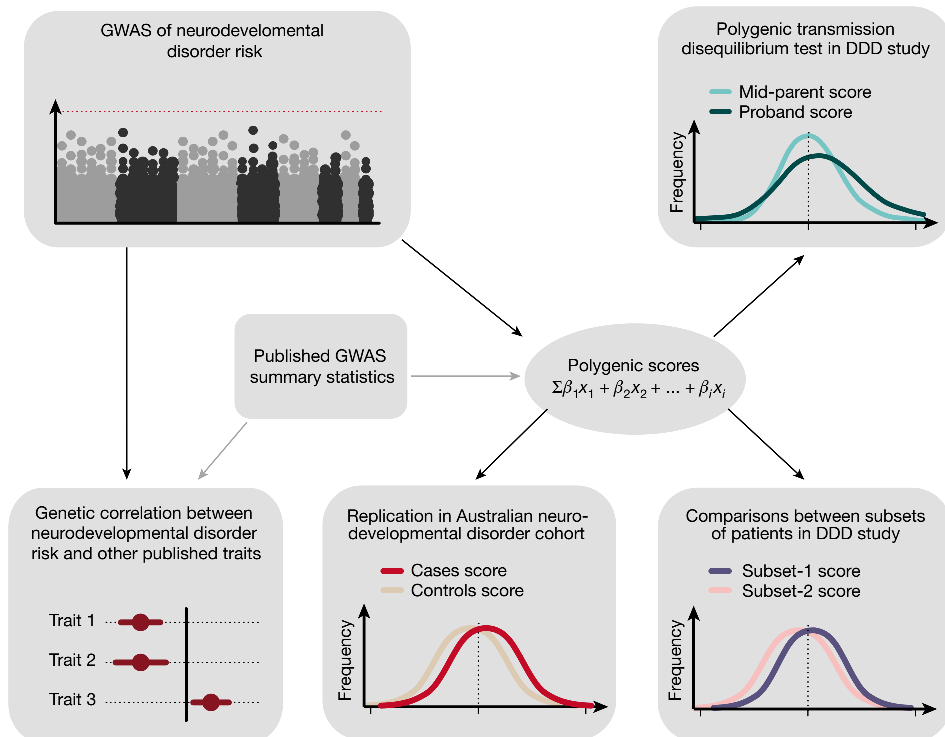


Fig. 1 | Outline of analysis exploring the contribution of common variants to risk of severe neurodevelopmental disorders. We first conducted a discovery GWAS in a large dataset of patients with neurodevelopmental disorders, and replicated the common-variant contribution by analysing polygenic transmission in independent trios from the same cohort. Next, we looked for overlap of common-variant effects between neurodevelopmental disorder risk and other published GWAS, and replicated these findings in an independent Australian cohort. Finally, we explored how polygenic effects were distributed within our discovery cohort of patients, and whether common variants contributed to expressivity of specific phenotypes.

in cells of the central nervous system ($P = 0.02$), and in mammalian constrained regions[18] ($P = 0.009$) (Supplementary Table 2), consistent with similar analyses for other neuropsychiatric and cognitive traits. Together, these results suggest that thousands of common variants have individually small effects on brain development or function, which in turn influences neuropsychiatric disease risk, cognitive traits and risk for severe neurodevelopmental disorders.

We next investigated how general our genetic correlation findings were by attempting to replicate them in another cohort of patients with neurodevelopmental disorders (Fig. 1). We obtained GWAS data for 1,270 neurodevelopmental disorder cases from Australia, and 1,688 ancestry-matched Australian controls. This sample size is too small to do direct genetic discovery or to reliably apply LD score regression, so we tested common-variant polygenic scores using summary statistics from our discovery GWAS and published GWAS, including educational attainment[15] and intelligence[16]. This approach requires specification of P-value thresholds and is less robust to population structure and cryptic relatedness, but it produced similar results to the genetic correlation analyses in our discovery GWAS and we therefore believe it is well-suited to a replication analysis. We replicated our observation of lower polygenic scores for educational attainment and intelligence in neurodevelopmental disorder cases from Australia, as compared to controls ($P = 1.0 \times 10^{-8}$ and $P = 7.6 \times 10^{-4}$ for educational attainment and intelligence, respectively), and found that cases had a nominally significantly increased score for schizophrenia ($P = 0.014$) (Methods and Extended Data Table 3). We did not see a significant difference between Australian cases and controls for the score constructed from our own discovery GWAS. If the two cohorts had identical phenotypes, we should have had 95% power (Methods) to detect a difference; this suggests that differences in how the British and Australian cohorts were recruited diluted our ability to quantify their shared genetics.

These findings could mean that common variants entirely explain a subset of patients with neurodevelopmental disorders and are not relevant in the remainder, or that the disorders of all patients have both rare- and common-variant contributions (Fig. 1). We have exome-sequenced our cohort of patients as well as their parents, and have previously reported a variety of both de novo and inherited diagnostic variants[19,20]. We therefore compared polygenic scores for cognitive traits and neuropsychiatric disorders between patients for whom we had identified diagnostic or probably diagnostic variants in a known developmental-disorder gene[21] ($n = 1{,}127$) and those who had no candidate diagnostic variant ($n = 2{,}479$), and found no significant differences for any polygenic score that we tested, after controlling for multiple testing (Extended Data Table 4 and Methods). We showed by simulations that if the 'diagnosed' cases had the same distribution of the polygenic score for educational attainment as did controls, we would have had sufficient power to detect a difference between them and the undiagnosed cases (Methods). This is consistent with a previous study in autism[11] that similarly found no evidence for a difference in polygenic risk scores between autism cases with a de novo diagnostic mutation compared to those without such a mutation. This suggests that in many patients both common and rare variants contribute to their neurodevelopmental disorder. However, as the DDD project continues to identify new diagnoses, we anticipate that the increase in power may show that monogenic and polygenic contributions are not purely additive.

In addition to showing that common variation affects overall risk of severe neurodevelopmental disorders, we sought to determine whether it can also affect individual presentation of symptoms. We identified four phenotypes measured in our neurodevelopmental disorder cohort for which independent GWAS data are available: autism (16% of the cohort), birth weight, height and intracranial volume. Compared to the age and sex-adjusted population average, our patients with neurodevelopmental disorders were—on average—0.72 s.d. shorter and weighed 0.15 s.d. less, and had a head circumference that was 1.20 s.d. smaller. We constructed common-variant polygenic scores for the four

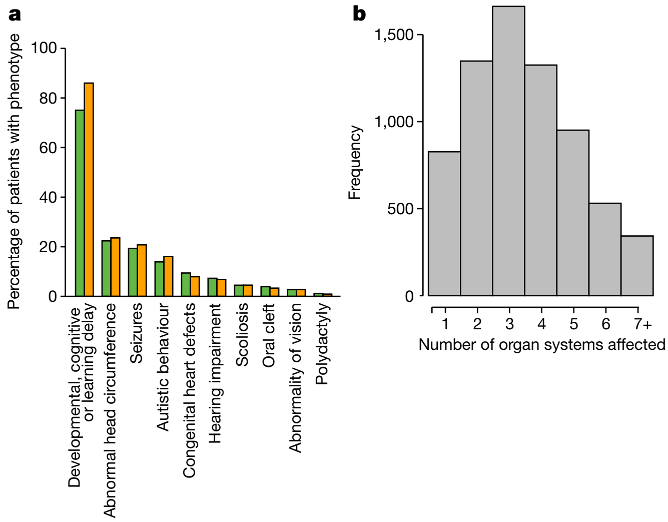


Fig. 2 | Patients recruited to the DDD study have diverse phenotypes. **a**, Examples of specific phenotypes that affect different organ systems, observed in the full DDD cohort ($n = 13{,}598$; green) and the subset of European patients with neurodevelopmental disorders ($n = 6{,}987$; orange). **b**, Distribution of the number of distinct organ systems that were affected in the set of 6,987 patients with neurodevelopmental abnormalities (Methods).

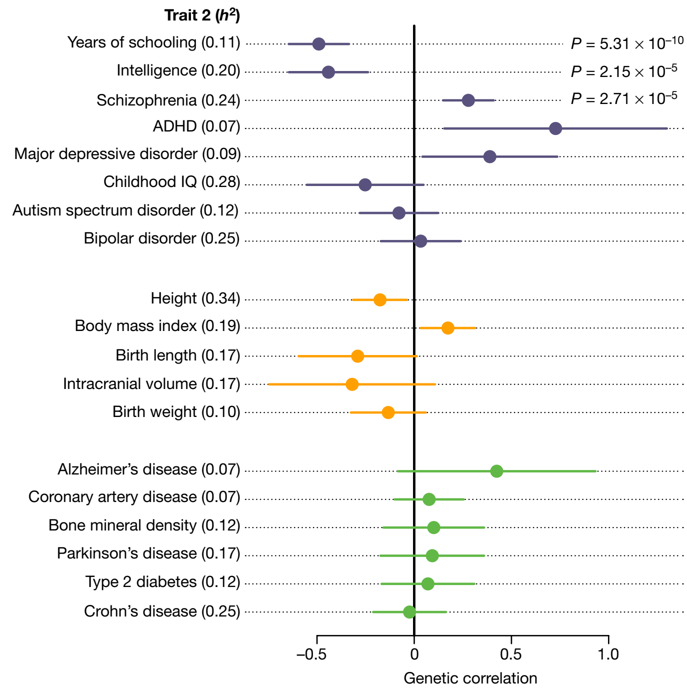


Fig. 3 | Genetic correlations between neurodevelopmental disorder risk (6,987 cases and 9,270 controls) against nineteen other traits. Cognitive or psychiatric (purple), anthropometric (orange) and negative-control traits (green), with single-nucleotide polymorphism (SNP) heritability (h^2) displayed for the trait. SNP heritability for dichotomous traits is displayed on the liability scale. Genetic correlation was calculated using bivariate LD score correlation[14], with the bars representing 95% confidence intervals (using standard error) before correction for multiple testing. Uncorrected P values are from a two-sided z-score, and are shown only if they pass Bonferroni correction for 19 traits. Sample sizes for 19 other GWAS are shown in Extended Data Table 2.

phenotypes as described above, and tested for an association between the relevant score and phenotype in our cohort. In all four cases, there was significant association (Table 1 and Extended Data Table 5), which demonstrates that common variation contributes to the expression of these traits in our study. Consistent with previous reports[9], we also found in our cohort that individuals with autism had higher polygenic scores for educational attainment compared to those without autism. We next tested for an association between the educational-attainment polygenic score and severity of the overall neurodevelopmental phenotype. We found that patients with severe intellectual disability or developmental delay ($n = 911$, Methods) had higher scores (that is, genetic predisposition to greater educational attainment; a proxy for higher cognitive function, $P = 0.004$, Table 1) than those with mild or moderate disability or delay ($n = 1{,}902$). This finding—which might seem initially counter-intuitive—is consistent with epidemiological studies[22] that have found that the siblings of patients with severe intellectual disability showed a normal distribution of IQ, whereas the siblings of patients with milder intellectual disability had lower IQ than average, which suggests that mild intellectual disability represents the tail-end of the distribution of polygenic effects on intelligence and severe intellectual disability has a different aetiology.

The study of human disease genetics has often been segregated into rare, single-gene disorders and common, complex disorders. There is abundant evidence that rare variants in individual genes can cause phenotypes that are seen much more commonly in individuals without a monogenic cause, including genes for maturity onset diabetes of the young[23] and familial Parkinson's disease[24]. There is also emerging evidence that the cumulative effect of common variants can modify the penetrance of rare variants in complex phenotypes such as educational attainment[25], schizophrenia[26] and breast cancer[27]. Here we have shown that the same interplay between rare and common variation exists even in severe neurodevelopmental disorders that are typically presumed to be monogenic. Previous studies have shown that the penetrance and expression of these disorders are affected by the specific missense variant that is carried[28] and the presence of mutations in secondary modifier genes[29]. Here we have demonstrated that phenotypic expression is also modified by common variants that influence neurodevelopmental traits in the general population. We analysed individuals of European ancestry—as do the vast majority of published GWAS—and, as the genetic architecture of neurodevelopmental disorders may differ

Table 1 | Polygenic score analyses in the DDD study

Measured trait	Polygenic score	Results[a] β	s.e.	P value	R^2
Birth weight ($n = 6{,}496$)	Birth weight	0.187	0.017	2.55×10^{-28}	0.020
Height ($n = 5{,}465$)	Height	0.408	0.033	1.18×10^{-35}	0.033
Head circumference ($n = 6{,}074$)	Intracranial volume	0.132	0.031	1.79×10^{-5}	0.004
Autistic behaviour: affected ($n = 1{,}121$), unaffected ($n = 5{,}866$)	Autism spectrum disorder	0.120	0.033	2.53×10^{-4}	0.006[c]
Developmental delay or intellectual disability: severe ($n = 911$), mild or moderate ($n = 1{,}902$)[b]	Educational attainment	0.116	0.040	0.004	0.008[c]

[a]Linear or logistic regression of measured traits in the DDD study against the respective polygenic score, including ten ancestry principal components as covariates. P values are two-sided, from t-distribution (linear) and z-score distribution (logistic).
[b]Severe cases were labelled as 1 in the logistic regression.
[c]Nagelkerke R^2.

between populations[20], further studies will be required to generalize our findings. Our findings suggest that fully understanding the genetic architecture of neurodevelopmental disorders will require considering the full spectrum of alleles, from those unique to an individual to those shared across continents.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0566-4.

Received: 4 May 2018; Accepted: 4 September 2018;
Published online 26 September 2018.

1. Boycott, K. M. et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am. J. Hum. Genet.* **100**, 695–705 (2017).
2. Owen, C. I. et al. Extending the phenotype associated with the *CSNK2A1*-related Okur–Chung syndrome—a clinical study of 11 individuals. *Am. J. Med. Genet. A* **176**, 1108–1114 (2018).
3. Singh, T. et al. Rare loss-of-function variants in *SETD1A* are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
4. Balasubramanian, M. et al. Delineating the phenotypic spectrum of Bainbridge-Ropers syndrome: 12 new patients with *de novo*, heterozygous, loss-of-function mutations in *ASXL3* and review of published literature. *J. Med. Genet.* **54**, 537–543 (2017).
5. Minikel, E. V. et al. Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).
6. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
7. The Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
8. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
9. Grove, J. et al. Common risk variants identified in autism spectrum disorder. Preprint at https://www.doi.org/10.1101/224774 (2017).
10. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
11. Weiner, D. J. et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* **49**, 978–985 (2017).
12. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
13. The Brainstorm Consortium. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
14. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
15. Lee, J.J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
16. Sniekers, S. et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
17. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
18. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
19. The Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
20. Martin, H. C. et al. Quantifying the contribution of recessive coding variation to developmental disorders. Preprint available at https://www.doi.org/10.1101/201533 (2017).
21. Firth, H. V. et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
22. Reichenberg, A. et al. Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proc. Natl Acad. Sci. USA* **113**, 1098–1103 (2016).
23. Flannick, J., Johansson, S. & Njølstad, P. R. Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nat. Rev. Endocrinol.* **12**, 394–406 (2016).
24. Hernandez, D. G., Reed, X. & Singleton, A. B. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. *J. Neurochem.* **139**, 59–74 (2016).
25. Ganna, A. et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19**, 1563–1565 (2016).
26. Tansey, K. E. et al. Common alleles contribute to schizophrenia in CNV carriers. *Mol. Psychiatry* **21**, 1085–1089 (2016).
27. Kuchenbaecker, K. B. et al. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in *BRCA1* and *BRCA2* mutation carriers. *J. Natl Cancer Inst.* **109**, djw302 (2017).
28. Martinelli, S. et al. Functional dysregulation of CDC42 causes diverse developmental phenotypes. *Am. J. Hum. Genet.* **102**, 309–329 (2018).
29. Khanna, H. et al. A common allele in *RPGRIP1L* is a modifier of retinal degeneration in ciliopathies. *Nat. Genet.* **41**, 739–745 (2009).

Acknowledgements We thank the families involved in the DDD study for their participation and patience, the DDD study clinicians, research nurses and clinical scientists in the recruiting centres for their hard work on behalf of families, M. Niemi for help making Fig. 1 and V. Warrier for useful discussions. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003), a parallel funding partnership between Wellcome and the Department of Health, and the Wellcome Sanger Institute (grant number WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of Wellcome or the Department of Health. The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. This study makes use of data generated by the DECIPHER community. Funding for the project was provided by the Wellcome Trust. We used data from 'Understanding Society: The UK Household Longitudinal Study', which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council (grant number ES/M008592/1). The data were collected by NatCen and the genome-wide scan data were analysed by the Wellcome Trust Sanger Institute. Data governance was provided by the METADAC data access committee, funded by ESRC, Wellcome and MRC (grant number MR/N01104X/1). Australian controls from the Brisbane Longitudinal Twin Study were collected and genotyped with grants from the National Health and Medical Research Council. We thank A. Pardiñas for producing the PGC-CLOZUK summary statistics without the Australian controls.

Reviewer information *Nature* thanks D. Arking, C. Lewis and S. Ripke for their contribution to the peer review of this work.

Author contributions J.C.B., C.F.W., D.R.F., H.V.F. and M.E.H. designed the study. M.E.K.N., H.C.M., D.L.R., G.G., M.K., J.M. and E.J.R. contributed to data analysis. S.Y., J.G. and N.G.M. performed data collection for the Australian cohort. K.M. and S.G. prepared data from the Australian cohort. M.E.K.N., H.C.M. and J.C.B. wrote the paper. H.C.M. and J.C.B. supervised the analyses and J.C.B. supervised the project.

Competing interests M.E.H. is a co-founder of, consultant to and holds shares in Congenica, a genetics diagnostic company. J.C.B. is an employee of Genomics plc.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0566-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0566-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.C.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Phenotypes of the DDD cohort. Recruitment and phenotyping of DDD patients is described in detail elsewhere[6,7]. The DDD study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). Families gave informed consent for participation. In brief, the DDD study recruited patients with a previously undiagnosed developmental disorder, in the UK and Ireland. Patient phenotypes were systematically recorded by clinical geneticists using Human Phenotype Ontology (HPO) terms in a central database, DECIPHER[21].

The DDD cohort is very heterogeneous in terms of patient phenotypes, and so we narrowed our analyses to singleton patients and trios where the proband had at least one of the following HPO terms (or daughter terms of these HPO terms): abnormal metabolic brain imaging by MRS (HP:0012705), abnormal brain positron emission tomography (HP:0012657), abnormal synaptic transmission (HP:0012535), abnormal nervous system electrophysiology (HP:0001311), behavioural abnormality (HP:0000708), seizures (HP:0001250), encephalopathy (HP:001298), abnormality of higher mental function (HP:0011446), neurodevelopmental abnormality (HP:0012759), abnormality of the nervous system morphology (HP:0012639). This 'neurodevelopmental' subset included individuals who have—since their recruitment to the DDD study—been found to carry diagnostic exome mutations in protein-coding genes[6,19,20,30], and individuals who are awaiting diagnosis. We therefore define our main phenotype ('neurodevelopmental disorder risk') as the risk of having a previously undiagnosed developmental disorder and being included in the DDD study, and having at least one neurodevelopmental HPO. In addition to HPOs, some DDD patients also had a clinical record of growth measurements such as height, birth weight and head circumference.

We counted the proportion of DDD patients with particular medically relevant HPOs, displayed in Fig. 2a. Individuals with the HPO were counted using a word search of the particular HPO and its daughter nodes. When counting the number of distinct organ systems affected in each DDD patient (Fig. 2b), we faced the issue that some HPOs fell under multiple organ systems: for example, microcephaly—which is a common term in the cohort—falls under three categories, 'nervous system', 'head or neck' and 'skeletal system'. To assign each HPO into only one organ system, we first ranked organ systems based on the number of raw counts of individuals with at least one term under that system (Extended Data Table 1) in the full DDD cohort. We then looked for individuals with at least one HPO under the organ system ranked most-commonly affected, and assigned these individuals an organ system count of 1. We then removed these HPOs from the lists of patients, before continuing to identify individuals with at least one HPO in the organ system ranked second-most prevalently affected. We continued to count organs and remove HPOs until we had assigned all individuals a count of organs systems affected out of 19 non-overlapping systems.

Developmental disorder phenotypes in the Australian cohort. We obtained a replication cohort of 1,270 cases of developmental disorder from South Australia, originally genotyped (using the Illumina Infinium CytoSNP-850k BeadChip) as part of routine clinical care to ascertain pathogenic copy-number variants. The majority (>95%) were under 18 years old. Between 50% and 60% were recruited through clinical genetics units, and the rest through neurologists, neonatologists, paediatricians and cardiologists. Based on reviewing information on the request forms, the majority of patients had developmental delay or intellectual disability, and malformations involving at least one organ (for example, brain, heart and kidney). Between 15% and 20% were recruited as neonates with multiple malformations involving brain, heart and/or other organs, and were too young to be diagnosed with developmental delay or intellectual disability.

Datasets and quality control. We genotyped 11,304 patients and 930 full trios recruited to the DDD study on Illumina HumanCoreExome and HumanOmniExpress chips, respectively. Genotyping was carried out by the Wellcome Trust Sanger Institute genotyping facility. As controls for the discovery GWAS, we used genotype data for 10,484 individuals from the UK-based 'Understanding Society' UK Household Longitudinal Study (UKHLS)[31,32]. Recruitment to this study was carried out through UK-wide household longitudinal survey. For replication, we obtained GWAS data from a cohort of cases of neurodevelopmental disorder from South Australia, and population-matched controls from the Brisbane Longitudinal Twin Study (Queensland Institute of Medical Research[33,34]). All data were on GRCh37, and detailed information of genotyping chips is shown in Supplementary Table 1.

We performed variant and sample quality control for each dataset separately. We removed samples of patients whose reported sex was inconsistent with the genotype data, who had high sample missingness ($\geq$3% of minor allele frequency (MAF) $\geq$ 10% variants), samples with high or low heterozygosity ($\pm$ 3 s.d. from the mean, using MAF $\geq$ 10% variants) to control for admixture and inbreeding, and sample duplicates (alleles identical by descent $\geq$ 98%, using MAF >10% variants). We removed one individual from pairs of related individuals (alleles identical by descent >12%, using PLINK) from the case–control cohorts. Individuals in the discovery cohort were not related to the independent DDD trios. We also removed trios with a high number of Mendelian errors (>2,000 errors). For variant quality control, we removed variants if they had high genotype missingness ($\geq$3%), Hardy–Weinberg equilibrium test $P < 1 \times 10^{-5}$, no strand information, if they were duplicates, if the alleles were discordant between case and control datasets, or if alleles and their frequency in Europeans were discordant with HRC v.1.1 imputation reference panel. We only included variants on chromosomes 1–22. For the HumanCoreExome data and the Australian data, we removed rare variants with MAF $\leq$ 0.5% before imputation. Post-imputation, we removed imputed variants with imputation quality score INFO $\leq$ 0.9 or high missingness ($\geq$5%).

We defined sample ancestry based on a projection principal component analysis (PCA) using PLINK with 1000 Genomes Phase 3 populations, using SNPs that overlapped between the datasets (DDD + UKHLS and Australian cases + controls separately) and the reference populations. For this, we used SNPs with a MAF $\geq$ 10%, excluded A/T and G/C SNPs, removed regions of extended linkage disequilibrium (including the HLA region), and thinned the SNPs by pruning those with pairwise $r^2 > 0.2$ in batches of 50 SNPs with sliding windows of 5 ('–indep-pairwise 50 5 0.2' in PLINK). This left 52,836 SNPs for the projection PCA with the DDD and UKHLS data, and 40,626 SNPs with the Australian data. For analyses described in this paper, we carried forward individuals of European ancestry, defined by selecting samples clustering around the 1000 Genomes Great British (GBR) samples in the PCA (Extended Data Figs. 1, 3). The distribution of ancestries was different between cases and controls, probably due to marked differences in ascertainment (for example, individuals from ancestries with high levels of consanguinity are more likely to be recruited to studies of rare genetic disorders). Because we tightly filtered based on PCA, these differences do not affect our results.

Phasing and imputation. After sample and variant quality control, we imputed European samples from all datasets to boost the coverage of the genome for association testing and to increase overlap of datasets genotyped on different chips. We used reference-based haplotype phasing and imputation. The discovery GWAS cohorts genotyped on the HumanCoreExome backbone were phased and imputed together using variants that intersected between the different versions of the chip. Trios were phased and imputed in a second batch because they were genotyped on a different chip. We phased and imputed the Australian GWAS data in a third batch, using variants that intersected between the CytoSNP-850K chip and the Illumina 610K chip. None of the analyses in our paper were directly across batches, so there is no bias introduced by this approach. We used the Sanger Institute Imputation Service[35] to carry out phasing (using Eagle2 (v.2.0.5)[36]) and imputation (using PBWT[37]) on the DDD discovery dataset, DDD trios dataset and Australian dataset, selecting the Haplotype Reference Consortium as the reference panel (release 1.1, chromosomes 1–22 and X)[35].

Discovery GWAS of neurodevelopmental disorder risk. We carried out a GWAS for neurodevelopmental disorder risk in the discovery neurodevelopmental set of 6,987 cases and 9,270 controls of European ancestry only, using BOLT linear mixed models[38] with sex as a covariate. We included in our analysis genotyped variants or high-confidence imputed variants (INFO $\geq$ 0.9) with a MAF of $\geq$ 5%.

SNP heritability. From the discovery GWAS summary statistics, we removed the MHC region (chromosome 6 region 26–34 Mb), and estimated trait heritability using LD score[8] in LD Hub[39]. Given the ascertainment of the DDD neurodevelopmental cases in this study, estimating the true population prevalence was not feasible. We therefore estimated SNP heritability for our discovery GWAS on the liability scale for a range of prevalence between 0.2% and 2%, and found that SNP heritability varies from 5.5% (s.e. = 1.5%) to 9.1% (s.e. = 2.5%). We report heritability assuming a prevalence of 1% in the population. Heritability on the observed scale in our discovery GWAS was 13.8% (s.e. = 3.7%).

Polygenic transmission disequilibrium test. We used the previously described polygenic transmission disequilibrium test (pTDT) method[11] to investigate transmission disequilibrium of effect alleles for traits within DDD trios, using imputed genotype data. In brief, the test compares the means of two polygenic score distributions: one comprising the scores of the probands, and the other the average parent-pair scores. The test is equivalent to a one-sample *t*-test, assessing whether the mean of score distribution in probands deviates from the mean of parent-pair score average. We report a one-sided *P* value for over-transmission.

Genetic correlation. We carried out genetic correlation of the neurodevelopmental disorder risk in the discovery GWAS against multiple published traits, using bivariate LD score[14]. For traits included in LD Hub, we used the online server, and for traits not included in LD Hub, we used the LD score software. For genetic correlation with neurodevelopmental disorder risk, we pre-selected a range of different types of traits and diseases: traits relating to cognitive performance, education, psychiatric traits and diseases, anthropometric traits and non-brain related traits

and diseases. Ninety-five per cent confidence intervals in Fig. 3 are shown before correction for multiple testing. We set the significance threshold to $P < 0.0026$ (0.05 divided by the 19 tests).

Partitioned heritability. We used partitioned LD score[17] to look for enrichment of heritability in cell-type groups and functional genomic categories. To do this, we used the baseline model LD scores and regression weights available online. For cell-type groups and functional categories, we set the significance threshold to $P < 0.005$ (0.05, divided by 10 tests) and $P < 9.6 \times 10^{-4}$ (0.05, divided by 52 tests), respectively.

Polygenic scores. We constructed polygenic scores using summary statistics from our GWAS of neurodevelopmental disorder risk and seven published GWAS (educational attainment[15], intelligence[16], schizophrenia[40], autism[9], intracranial volume[41], height[42] and birth weight[43]). For all traits, we included only variants that had a MAF $\geq$ 5% and were directly genotyped or imputed with high confidence (INFO ≥ 0.9) in the respective study cohort (discovery case and control, trios or Australian). To construct the polygenic scores for individuals, we then multiplied the variant effects (β values) with the allele counts of each individual. For imputed variants, we used genotype probabilities rather than hard-called allele counts. To find independent variants for our scores, we pruned variants intersecting the original study summary statistics and our GWAS data using PLINK, by taking the top variant and removing variants within 500 kb and that have $r^2 \geq 0.1$ with the top variant. We then repeated the process until no variant had a P value below a predefined threshold, which we based on prior knowledge of variance in the phenotype explained. For the neurodevelopmental disorder risk score, we tested seven P-value thresholds ($P < 1$, 0.5, 0.1, 0.05, 0.01, 0.005 and 0.001) and chose the one which resulted in a score that explained the most variance (Nagelkerke's R^2) in case and control status in an independent subset of DDD patients. Specifically, we repeated our GWAS of neurodevelopmental disorder risk having removed a random subset of 20% of cases and controls, then calculated a score in this 'leave-out' subset, and performed a logistic regression to assess association of case–control status with the score. The threshold $P < 1$ performed best in ten independent permutations, and we used this threshold to construct scores in pTDT and Australian case–control analyses. We additionally tested all seven thresholds when constructing scores in the Australian dataset; however, varying the threshold did not change our results. When deciding the P-value thresholds for published GWAS, we used the threshold that had been found to explain the most variation in other published studies for the trait (years in education[44] $P < 1$, intelligence[16] and schizophrenia[40] $P < 0.05$, and autism[11] $P < 0.1$). For traits for which we had phenotype data in the DDD, we used thresholds that explained the most variation in DDD cases (intracranial volume $P < 1$, birth weight $P < 0.01$ and height $P < 0.005$). Thresholds and the number of variants used for each score are shown in Extended Data Tables 3–5. All scores were normalized to a mean of 0 and variance of 1. To test for association between trait and score, we used R (version 1.90b3) to perform logistic regression for binary traits and linear regression for quantitative traits, including the first ten principal components from the ancestry PCA to control for possible population stratification.

To assess power for detecting differences in scores between diagnosed and undiagnosed patients, we tested the hypothesis that diagnosed patients were effectively a random sample of controls with respect to their polygenic scores. Specifically, we randomly sampled 1,127 controls (that is, the same number as we had diagnosed patients) and compared the polygenic scores between them and the undiagnosed patients using logistic regression. We repeated this 10,000 times and determined the proportion of times we detected a significant difference $P < 0.007$ ($P < 0.05$, divided by 7 (correcting for seven polygenic scores)) as proxy for power. This was 99.1% of simulations for educational attainment, 93.6% of simulations for schizophrenia and 61.2% of simulations for intelligence.

We used AVENGEME[45] to calculate power to find significant association (at $P < 0.05$) between our polygenic score for neurodevelopmental disorders and case or control status in the Australian dataset. We assumed that the SNP heritability is the same (7.7%) in both the Australian and British cohorts, and that the genetic correlation between them was 1.

The PGC-CLOZUK study of schizophrenia included some controls from the Australian cohort used in our study, and therefore we ran polygenic score analyses in the Australians using summary statistics from PGC-CLOZUK (obtained from A. Pardiñas, personal communication) after these samples had been removed.

Producing subsets from the DDD cohort. We defined a set of patients with an exonic diagnosis and a set with no likely diagnostic variants. This was based on a previously described clinical filtering procedure[6]—which focuses on identifying rare, damaging variants in a set of genes known to cause developmental disorders (https://www.ebi.ac.uk/gene2phenotype/)—that fit an appropriate inheritance mode. Variants that pass clinical filtering are uploaded to DECIPHER, where the patients' clinicians classify them as 'definitely pathogenic', 'likely pathogenic', 'uncertain', 'likely benign' or 'benign'. This process of clinical classification is necessarily dynamic as new disorders are identified and patients manifest new phenotypes. Our 'diagnosed' set consists of 1,127 patients who fulfilled one of the following criteria: (a) among the diagnosed set in a recent reanalysis of the first 1,133 trios[46]; (b) had at least one variant (or pair of compound heterozygous variants) rated as 'definitely pathogenic' or 'likely pathogenic' by a clinician; or (c) had at least one variant (or pair of compound heterozygous variants) in a class with a high positive-predictive value that passed clinical filtering but had not yet been rated by clinicians. We considered de novo or compound heterozygous loss-of-function (LOF) variants to have high positive-predictive value, as of the ones that had been rated by clinicians, 100% of compound heterozygous LOFs and 94.% of de novo LOFs had been classed as 'definitely pathogenic' or 'likely pathogenic'. Our 'undiagnosed' set consists of 2,479 patients who had no variants that passed our clinical filtering, or in whom the variants that had passed clinical filtering had all been rated as 'likely benign' or 'benign' by clinicians, or who were among the undiagnosed set in the first 1,133 trios that have previously been extensively clinically reviewed[6]. Note that our diagnosed versus undiagnosed analysis excludes 3,375 patients who had one or more variants that passed clinical filtering in a class with a relatively low positive-predictive value, but that have not yet been rated by clinicians.

We defined patients to present with autistic behaviours if their phenotype included autistic behaviour (HP:0000729) or any of its daughter nodes. We defined patients as having 'mild/moderate intellectual disability or delay' if their HPO phenotypes included borderline, mild or moderate intellectual disability (HP:0006889, HP:0001256, HP:0002342) and/or mild or moderate global developmental delay (HP:0011342, HP:0011343). Patients were included in the 'severe ID or delay' set if they had severe or profound intellectual disability (HP:0010864, HP:0002187) and/or severe or profound global developmental delay (HP:0011344, HP:0012736). We excluded patients with intellectual disability or global developmental delay of undefined severity.

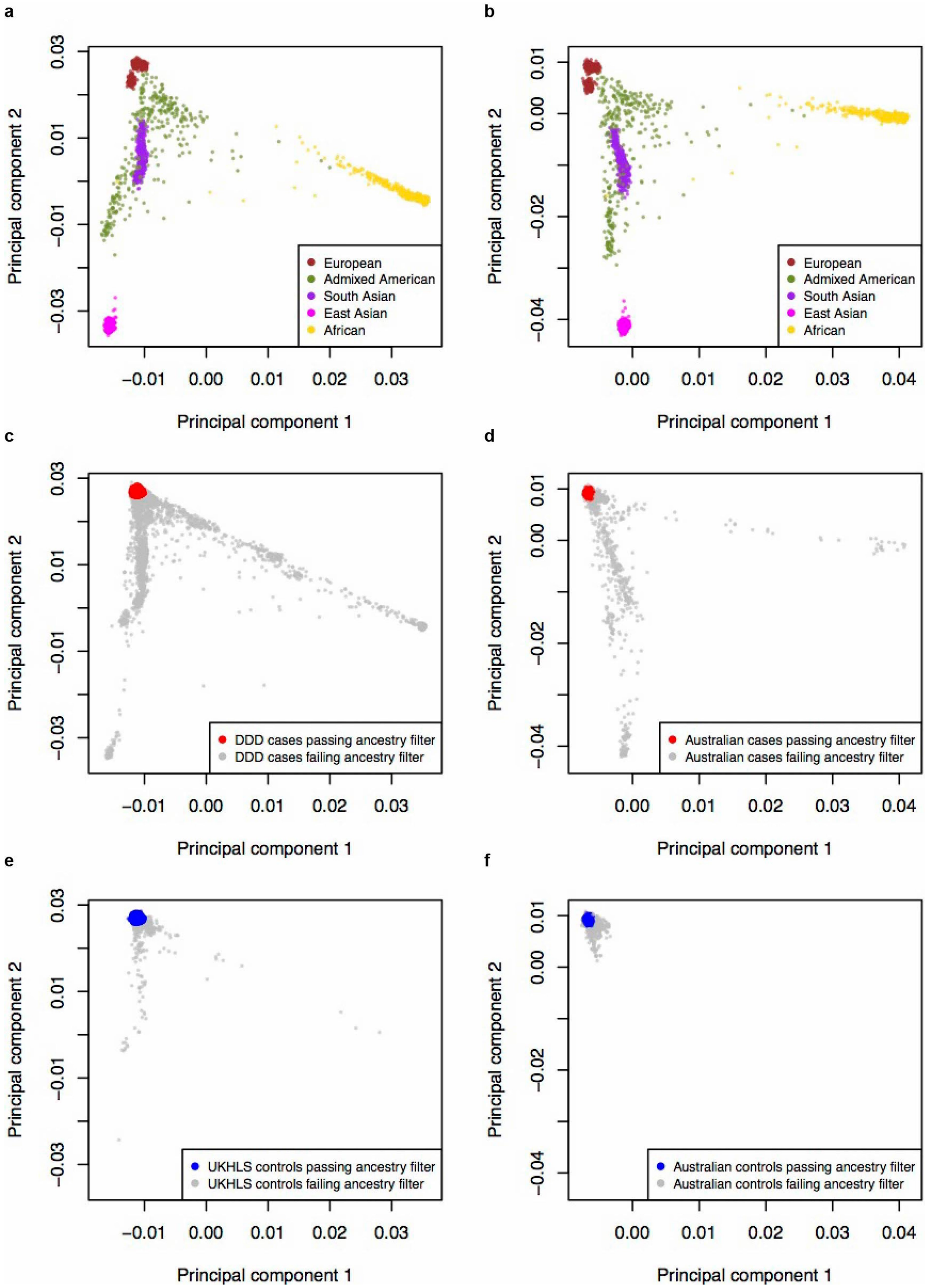
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

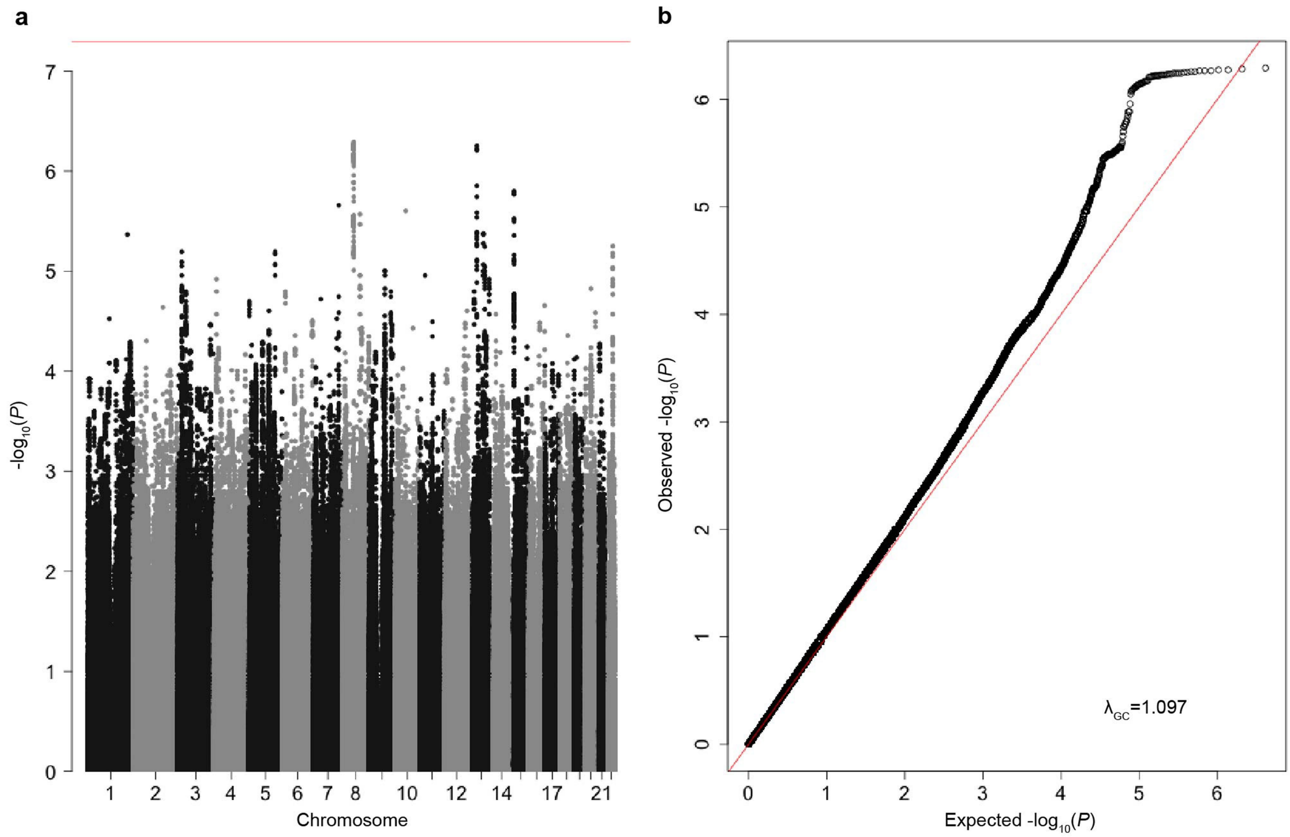
The raw genotype data, post-quality-control genotype data and discovery GWAS summary statistics generated and/or analysed during the current study are available through European Genome-phenome Archive, under EGAS00001000775. This study makes use of data generated by the DECIPHER community: a full list of centres that contributed to the generation of the data is available from http://decipher.sanger.ac.uk, and via email from decipher@sanger.ac.uk. Information on how to access the data from the UKHLS can be found on the 'Understanding Society' website, at https://www.understandingsociety.ac.uk/.

30. Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
31. University of Essex Institute for Social and Economic Research, NatCen Social Research & Kantar Public. *Understanding Society: Waves 1–7, 2009–2016 and Harmonised BHPS: Waves 1–18, 1991–2009* 10th edn (UK Data Service, 2018).
32. University of Essex Institute for Social and Economic Research. *Understanding Society: Waves 2 and 3 Nurse Health Assessment, 2010–2012* 3rd edn (UK Data Service, 2014).
33. Wright, M. J. & Martin, N. G. Brisbane adolescent twin study: outline of study methods and research projects. *Aust. J. Psychol.* **56**, 65–78 (2004).
34. Mina-Vargas, A. et al. Heritability and GWAS analyses of acne in Australian adolescent twins. *Twin Res. Hum. Genet.* **20**, 541–549 (2017).
35. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
36. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
37. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
38. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
39. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
40. Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
41. Adams, H. H. H. et al. Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat. Neurosci.* **19**, 1569–1582 (2016).
42. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
43. Horikoshi, M. et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).
44. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).

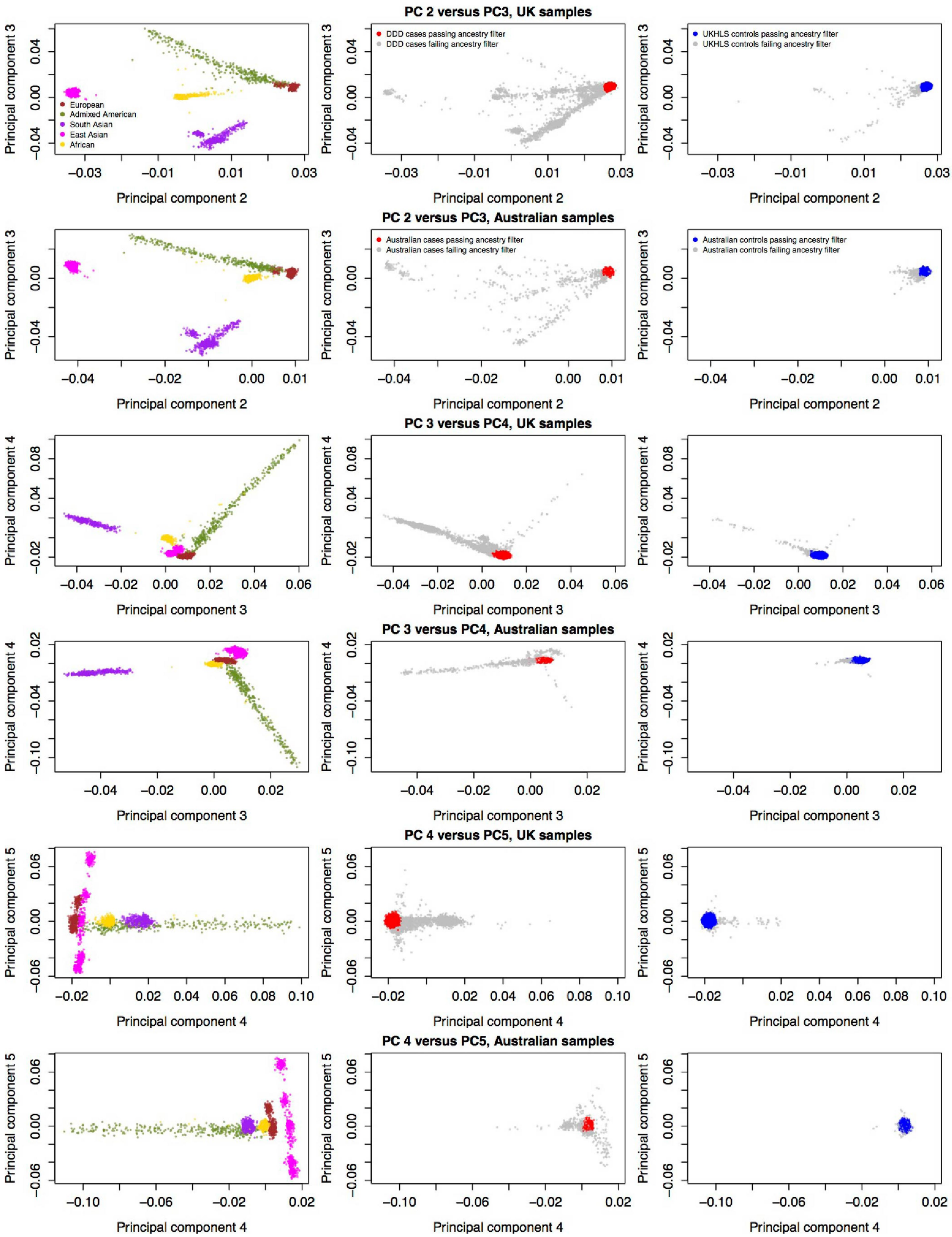
45. Palla, L. & Dudbridge, F. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am. J. Hum. Genet.* **97**, 250–259 (2015).
46. Wright, C. F. et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med*. https://doi.org/10.1038/gim.2017.246 (2018).



Extended Data Fig. 1 | Ancestry principal components analysis of UK and Australian samples. a, **b**, Reference samples ($n = 2{,}504$) from 1000 Genomes Phase 3, coloured by the five super-populations, used for a projection PCA of UK cohorts (DDD and UKHLS) (**a**) or Australian cohorts (**b**). **c**, **d**, All DDD cases (discovery $n = 11{,}304$ and from trios $n = 930$) (**c**) and all Australian cases ($n = 2{,}283$) (**d**) from their respective projection PCA with 1000 Genomes. Case samples from individuals with European ancestry are plotted in red and non-Europeans in grey. **e**, **f**, All UKHLS controls ($n = 10{,}396$) (**e**) and all Australian controls ($n = 4{,}274$) (**f**) from their respective projection PCA with 1000 Genomes. Control samples from individuals with European ancestry are plotted in blue and non-Europeans in grey. All cases and controls coloured in grey (in **c**–**f**) were excluded from analysis owing to non-European ancestry. UK cohorts are plotted after removal of samples that failed quality control, and Australian cohorts before removal of samples that failed quality control.



Extended Data Fig. 2 | Discovery GWAS of neurodevelopmental disorder risk. a, Manhattan plot of discovery GWAS of neurodevelopmental disorder risk, with 6,987 DDD cases and 9,270 ancestry-matched UKHLS controls (both for individuals with European ancestry), using 4,134,438 variants, MAF $\geq$ 5%, chromosomes 1–22. *P* values were from a two-tailed χ^2 distribution. Red line represents the threshold for genome-wide significance ($P = 5 \times 10^{-8}$). **b**, Quantile–quantile plot of discovery GWAS of neurodevelopmental disorder risk. Red line represents the expected values under the null hypothesis.



Extended Data Fig. 3 | Ancestry principal components analysis of samples from the UK and Australian (principal components 2–5). Reference samples ($n = 2{,}504$) from 1000 Genomes Phase 3—coloured by the five super-populations—are plotted on the left hand side, from projection PCAs with UK cohorts. Middle panels show the principal components plotted for DDD cases (discovery $n = 10{,}556$ and from trios $n = 911$) (UK samples) and Australian cases ($n = 2{,}283$). Red, case samples from individuals with European ancestry. Grey, samples from individuals of non-European ancestry; these individuals were excluded from analyses. Right-hand panels show principal components for UKHLS controls ($n = 10{,}396$) (UK samples) and Australian controls ($n = 4{,}274$). Blue, control samples from individuals with European ancestry. Grey, samples from individuals of non-European ancestry; these individuals were excluded from analyses. UK cohorts are plotted after removal of samples that failed quality control, and Australian cohorts before removal of samples that failed quality control.

Extended Data Table 1 | Proportions of patients with a neurodevelopmental disorder who have at least one HPO term that belongs to a particular organ-system category

Organ system	% All DDD patients (N=13,558)	% unrelated DDD patients, GBR ancestry (N=6,987)
Nervous system	87.0	100.0
Head or neck	68.9	71.2
Skeletal system	61.7	61.8
Limbs	35.1	35.3
Eye	34.9	35.3
Integument	31.2	31.9
Ear	20.1	19.7
Digestive system	20.0	19.1
Musculature	19.9	18.7
Cardiovascular system	15.1	13.5
Genitourinary system	12.4	11.4
Respiratory system	8.1	7.3
Connective tissue	7.4	6.3
Immune system	6.8	6.5
Endocrine system	4.1	4.1
Metabolism homeostasis	4.1	4.0
Breast	3.7	3.7
Blood and blood forming tissues	2.1	2.1
Voice	1.1	1.1

The HPO tree descends from ‘phenotypic abnormality’ through different organ systems down to specific terms that describe particular phenotypes. Each HPO term used by clinicians to describe patients was traced up the tree to the organ-system level. However, some HPOs may belong to more than one organ-system category: for example, microcephaly will be counted under ‘nervous system’, ‘head or neck’ and ‘skeletal system’ in the HPO tree whereas global developmental delay will appear only under ‘nervous system’.

Extended Data Table 2 | Genetic correlations between neurodevelopmental disorder risk and a range of traits, calculated using the LD score method

Trait 2	Genetic correlation between developmental disorder risk and trait 2	Standard error	95% confidence interval (standard error) lower bound	95% confidence interval (standard error) upper bound	P-value	SNP heritability for trait 2[a]	SE for trait 2 SNP heritability	Population prevalence used for liability scale conversion	N for trait 2 GWAS (Ncases:Ncontrols if dichotomous)
Years of schooling	-0.491	0.079	-0.336	-0.645	5.31×10^{-10}	0.112	0.004		766,345
Intelligence (Spearman's g)	-0.441	0.104	-0.237	-0.645	2.15×10^{-5}	0.203	0.013		78,308
Schizophrenia	0.279	0.066	0.148	0.409	2.71×10^{-5}	0.242	0.008	0.010	11,260:24,542
Attention deficit hyperactivity disorder	0.727	0.292	0.155	1.299	0.013	0.071	0.031		17,666
Major depressive disorder	0.389	0.177	0.042	0.736	0.028	0.087	0.017	0.150	9,240:9,519
Childhood IQ	-0.252	0.153	0.048	-0.553	0.100	0.279	0.051		12,441
Autism spectrum disorder	-0.078	0.103	0.123	-0.280	0.445	0.118	0.010	0.012	18,381:27,969
Bipolar disorder	0.033	0.105	-0.172	0.238	0.751	0.250	0.023	0.010	7,481:9,250
Height	-0.176	0.070	-0.038	-0.314	0.012	0.336	0.021		253,288
Body mass index	0.174	0.071	0.035	0.312	0.015	0.189	0.010		336,107
Child birth length	-0.291	0.155	0.013	-0.595	0.061	0.165	0.027		28,459
Intracranial volume	-0.319	0.218	0.107	-0.746	0.142	0.167	0.053		11,373
Birth weight	-0.133	0.098	0.059	-0.326	0.174	0.095	0.008		143,677
Alzheimer's disease	0.424	0.259	-0.083	0.932	0.101	0.068	0.013	0.050	17,008:37,154
Coronary artery disease	0.077	0.091	-0.101	0.254	0.396	0.070	0.005	0.050	60,801:123,504
Lumbar Spine bone mineral density	0.101	0.132	-0.158	0.360	0.447	0.116	0.018		44,731
Parkinson's disease	0.093	0.136	-0.173	0.359	0.494	0.167	0.050	0.002	1,713:3,978
Type 2 Diabetes	0.071	0.122	-0.168	0.309	0.562	0.120	0.147	0.015	12,171:56,862
Crohn's disease	-0.024	0.096	0.164	-0.211	0.804	0.252	0.027	0.003	5,956:14,927

Trait 2 is the trait to which neurodevelopmental disorder risk is compared. Uncorrected *P* values are from a two-sided *z*-score.
[a]SNP heritability for dichotomous traits is on the liability scale.

Extended Data Table 3 | Polygenic score analyses comparing 1,266 Australian cases of neurodevelopmental disorders and 1,688 controls

	Polygenic score parameters			Results[a]		
Polygenic score	r^2 for SNP pruning	P-value threshold for SNP pruning	Number of SNPs in score	Beta	Standard error	P-value
Educational attainment (SSGAC, 2018)	0.1	1	92,091	-0.218	0.038	9.97×10^{-9}
Height (Wood et al., 2014)	0.1	0.005	9,809	-0.155	0.040	8.84×10^{-5}
Intelligence (Sniekers et al., 2017)	0.1	0.05	21,551	-0.126	0.038	7.61×10^{-4}
Schizophrenia (QIMR removed) (Pardinas et al., 2018)	0.1	0.05	23,878	0.092	0.038	0.014
Intracranial volume (Adams et al., 2016)	0.1	1	90,928	-0.078	0.038	0.041
Autism (Grove et al., 2017)	0.1	0.1	26,846	0.070	0.038	0.063
Birth weight (Horikoshi et al., 2016)	0.1	0.01	6,828	-0.062	0.038	0.098
Developmental disorder risk (discovery GWAS)	0.1	1	67,001	-0.047	0.038	0.212

P values are uncorrected, two-sided and from *z*-score distribution. Data were obtained from previous studies[9,15,16,40–43].
[a]Logistic regression of case or control status on polygenic score, using ten ancestry principal components as covariates.

Extended Data Table 4 | Polygenic score analyses comparing patients from the DDD with an exome diagnosis (*n* = 1,127) against undiagnosed patients (*n* = 2,479)

	Parameters			Results[a]		
Polygenic score trait	r^2 for SNP pruning	P-value threshold for SNP pruning	Number of SNPs in score	Beta	Standard error	P-value
Educational attainment	0.1	1	79,292	0.080	0.037	0.028
Intelligence	0.1	0.05	19,387	0.063	0.036	0.080
Schizophrenia	0.1	0.05	21,321	0.017	0.036	0.644
Autism	0.1	0.1	23,648	-0.077	0.036	0.032
Intracranial volume	0.1	1	76,788	4.98×10^{-3}	0.036	0.891
Birth weight	0.1	0.01	6,212	1.54×10^{-3}	0.036	0.966
Height	0.1	0.005	9,019	1.34×10^{-3}	0.036	0.971

P values are uncorrected, two-sided and from *z*-score distribution.
[a]Logistic regression of diagnosed and undiagnosed status on polygenic score, using ten ancestry principal components as covariates.

Extended Data Table 5 | Polygenic score analyses in patients from the DDD for measured traits

		Score parameters			Results[a]			
Measured trait	Polygenic score	r^2 for SNP pruning	P-value threshold for SNP pruning	Number of SNPs in score	Beta	Standard error	P-value	R^2
Birth weight (N=6,496)	Birth weight	0.1	0.01	6,212	0.187	0.017	2.55×10^{-28}	0.020
Height (N=5,465)	Height	0.1	0.005	9,019	0.408	0.033	1.18×10^{-35}	0.033
Head circumference (N=6,074)	Intracranial volume	0.1	1	76,788	0.132	0.031	1.79×10^{-5}	0.004
Autistic behavior: affected (N=1,121), unaffected (N=5,866)	Autism	0.1	0.1	23,648	0.120	0.033	2.53×10^{-4}	0.006[c]
Developmental delay or intellectual disability: severe (N=911), mild/moderate (N=1,902)[b]	Educational attainment	0.1	1	79,292	0.116	0.040	0.004	0.008[c]

P values are uncorrected, two-sided and from *t*-distribution (linear) and *z*-score distribution (logistic). [a]Linear or logistic regression on polygenic score. using ten ancestry principal components as covariates.
[b]Severe cases were labelled as 1 in the logistic regression.
[c]Nagelkerke R^2.

LETTER

https://doi.org/10.1038/s41586-018-0578-0

Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids

Leonardo Beccari[1,6], Naomi Moris[2,6], Mehmet Girgin[3,6], David A. Turner[2], Peter Baillie-Johnson[2,5], Anne-Catherine Cossy[4], Matthias P. Lutolf[3], Denis Duboule[1,4,7]* & Alfonso Martinez Arias[2,7]*

The emergence of multiple axes is an essential element in the establishment of the mammalian body plan. This process takes place shortly after implantation of the embryo within the uterus and relies on the activity of gene regulatory networks that coordinate transcription in space and time. Whereas genetic approaches have revealed important aspects of these processes[1], a mechanistic understanding is hampered by the poor experimental accessibility of early post-implantation stages. Here we show that small aggregates of mouse embryonic stem cells (ESCs), when stimulated to undergo gastrulation-like events and elongation in vitro, can organize a post-occipital pattern of neural, mesodermal and endodermal derivatives that mimic embryonic spatial and temporal gene expression. The establishment of the three major body axes in these 'gastruloids'[2,3] suggests that the mechanisms involved are interdependent. Specifically, gastruloids display the hallmarks of axial gene regulatory systems as exemplified by the implementation of collinear *Hox* transcriptional patterns along an extending antero-posterior axis. These results reveal an unanticipated self-organizing capacity of aggregated ESCs and suggest that gastruloids could be used as a complementary system to study early developmental events in the mammalian embryo.

Recent work on organoids derived from stem cells has revealed a surprising autonomy in the development of particular tissues and organs[4,5]. When around 250 ESCs are aggregated, given a pulse of the Wnt agonist CHIR99021 (Chi) between 48 and 72 h after the start of culture, and returned to N2B27 medium (Fig. 1a), a pole of *Bra* (brachyury, also known as *T*) expression emerges reproducibly[6] (Fig. 1b, Extended Data Fig. 1), resembling the elongating embryonic tail bud. The aggregates continue to elongate up to 120 h after aggregation (AA), when they display a 'rostral' cell-dense region and a polar extension towards a 'caudal' extremity, reaching up to 500 μm in size (Fig. 1b). Shaking the culture enables aggregates to grow to 850–1,000 μm in length at 168 h AA (Fig. 1c, d). At 120 h AA, a *Gata6*-positive domain is visible opposite to a *Bra* and *Cdx2*-expressing region, probably corresponding to the cardiac crescent, which delimits the embryonic post-occipital region[7] (Fig. 1b–d, Extended Data Fig. 1, Supplementary Videos 1, 2). By contrast, at 120–168 h AA *Sox1*/*Sox2*-positive cells are localized centrally, with the exception of those at the rostral extremity (Fig. 1c, d).

To characterize the transcriptional programmes of these gastruloids, we carried out RNA-sequencing (RNA-seq) analysis on duplicate pools of gastruloids and compared their profiles with those of developing mouse embryos from E6.5 to E9.5. Because gastruloids display

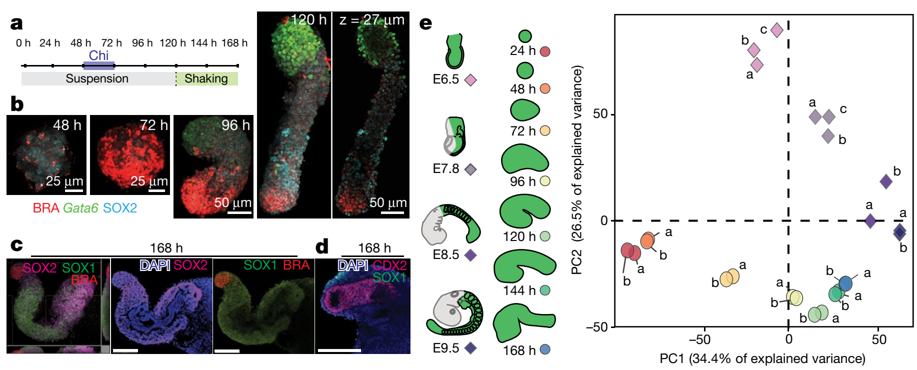


Fig. 1 | Elongation of gastruloids. a, Schematic of the culture protocol: 200–300 ESCs were allowed to aggregate; the *Wnt* agonist CHIR99201 (Chi) was added between 48 h and 72 h AA; organoids were cultured in suspension until 120 h AA (grey rectangle) and transferred into shaking cultures until 168 h AA. **b**, Three-dimensional renderings and confocal sections of gastruloids at different times showing the elongation and expression of BRA, SOX2 and $Gata6^{H2B\text{-}Venus}$ (green). The right-most panel is a confocal section of the 3D rendering of the neighbouring 120 h AA gastruloid. Scale bars, 25 μm (48 h, 72 h), 50 μm (96 h, 120 h). Each image is representative of an experiment with seven biological replicates showing the same expression pattern. **c**, **d**, Three-dimensional rendering (**c**, left) and confocal sections (**c**, centre, right and **d**, tail region) of gastruloids at 168 h AA, showing the localization of CDX2, SOX2, SOX1 and BRA proteins. Scale bar, 150 μm. Each image is representative of an experiment performed in 20 biological replicates. The reported expression pattern was observed in at least 80% of the cases. **e**, PCA analysis of RNA-seq datasets using time-pooled gastruloids from 24 to 168 h AA ($n = 2$ replicates per time point) and pooled mouse embryos at E6.5 ($n = 3$), E7.8 ($n = 3$), E8.5 (12–14 somites, $n = 2$ replicates) and E9.5 (~24 somites, $n = 2$ replicates). Each replicate was derived from an independent sample. For E7.8 embryos, only their posterior half was used. For E8.5 and E9.5, the post-occipital embryonic domain was dissected. In all cases, the portion used for RNA extraction is coloured in pale green. All autosomal genes were considered for this analysis. PC1 shows a strong temporal component whereas PC2 discriminates between gastruloids or embryonic samples.

[1]Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland. [2]Department of Genetics, University of Cambridge, Cambridge, UK. [3]School of Life Sciences, Federal Institute of Technology EPFL, Lausanne, Switzerland. [4]Swiss Cancer Research Institute (ISREC), School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [5]Present address: Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [6]These authors contributed equally: Leonardo Beccari, Naomi Moris, Mehmet Girgin. [7]These authors jointly supervised this work: Denis Duboule, Alfonso Martinez Arias. *e-mail: denis.duboule@epfl.ch; ama11@hermes.cam.ac.uk

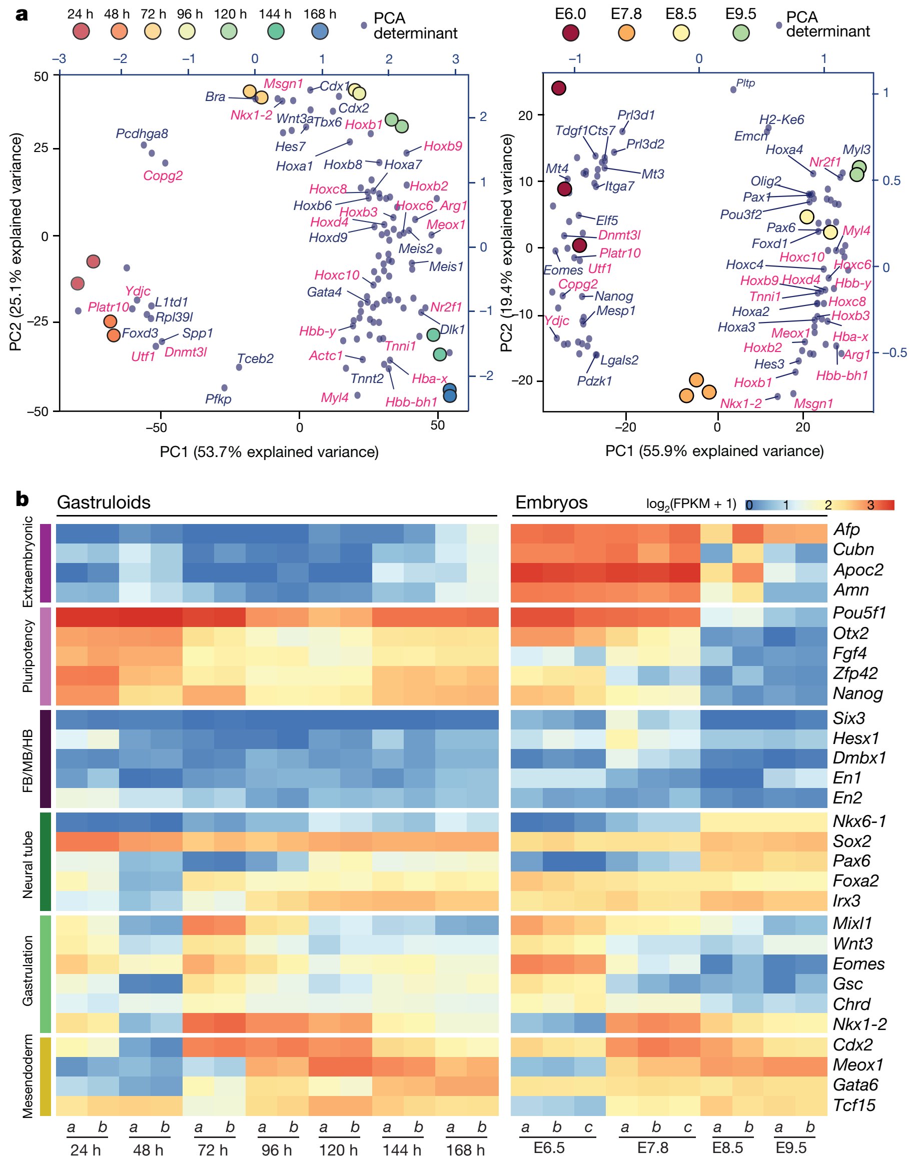


Fig. 2 | Temporal patterns of gene expression in gastruloids. a, PCA of either pooled gastruloids during temporal progression from 24 h to 168 h AA (left) or mouse embryos from E6.5 to E9.5 (right). The 100 top contributing genes to the first two principal components are labelled, with those common to both gastruloid and embryo datasets in red. *Tceb2* is also known as *Elob*. **b**, Heat map of scaled expression of genes associated with development of different embryonic structures in pooled gastruloids and embryos over time. The replicates represented in these graphs were derived from biologically independent samples, as in Fig. 1e. FB, forebrain; MB, midbrain; HB, hindbrain.

hallmarks of post-occipital embryos[6] (Fig. 1b–d) we excluded the anterior portion of E7.5–E9.5 embryos (Fig. 1e, left). Principal component analysis (PCA) demonstrated reproducibility between samples and a clear clustering along principal component 1 (PC1) corresponding to their temporal order (Fig. 1e), whereas embryo samples segregated from gastruloids along PC2 only. The main (top 100) clustering determinants of gastruloid samples included several pluripotency-related genes, epiblast markers and genes involved in gastrulation. They also comprised different *Hox* genes and other transcription factors usually expressed in post-occipital structures of the developing mouse embryo such as *Cdx1*, *Cdx2*, *Meis1*, *Meis2*, *Meox1*, *Bra* and *Gata4* (Fig. 2a). Twenty-five of these 100 PCA determinants were identified independently in both gastruloid and embryonic temporal series (Fig. 2a, genes in red), supporting the idea that gastruloids and embryos elongate by implementing similar transcriptional programs. The analysis of specific genes associated with particular developmental landmarks confirmed this point (Fig. 2b, Extended Data Fig. 2a, b). For instance, genes associated with gastrulation, such as *Mixl1*, *Eomes*, *Gsc* (goosecoid) or *Chrd* (chordin) were transiently and orderly transcribed at around 48 h AA (Fig. 2b, Extended Data Fig. 2b), suggesting that at this stage the gastruloid transcriptome resembles that of mouse epiblast at the onset of gastrulation. By 72 h AA, we observed an increase in the complexity of gene-expression profiles, with the appearance of markers for different embryonic lineages, including mesendoderm and neuroectoderm, and the transcription of *Hox* gene clusters (Fig. 2a, b, Extended Data Fig. 2a, b; see below). Genes associated with either extra-embryonic structures or anterior neural plate derivatives were not (or were poorly) expressed in gastruloids (Fig. 2b, Supplementary Information datasets 1, 2).

PCA analysis using single gastruloids revealed robust clustering at all assessed developmental stages and a correspondence with the pooled RNA-seq datasets (Extended Data Fig. 2c), showing that the population of gastruloids was comparatively homogenous at the time points analysed, and hence that the pooled RNA-seq datasets reflected the transcriptional status of single gastruloids. Transcriptome analyses of gastruloids revealed mRNAs that are usually associated with neural, endodermal and mesodermal derivatives, including paraxial, cardiac, intermediate and haematopoietic progenitors as well as neural

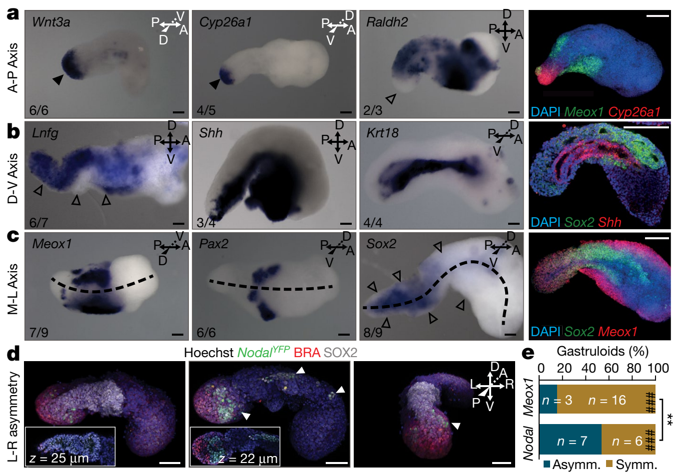


Fig. 3 | Multi-axial organization of gastruloids. a–c, Gene expression in gastruloids at 144 h AA, showing their axial organization. **a**, *Wnt3a* and *Cyp26a1* expression (arrowhead) at the posterior end, where *Raldh2* is not transcribed (empty arrowhead). Double-fluorescence in situ hybridization (FISH) staining of *Meox1* and *Cyp26a1* (**a**, far right). **b**, Dorso-ventral axis revealed by the ventral expression of *Shh* and *Krt18*, and of *Lfng* dorsally (empty arrowheads mark the ventral *Lnfg*-negative domain). Double FISH staining of *Sox2* and *Shh* confirms a dorso-ventral segregation, with *Shh* expressed exclusively in endoderm precursors (**b**, far right). **c**, Medio-lateral axis of symmetry (dotted line) revealed by the bilateral expression of *Meox1* and *Pax2*, complementary to the central distribution of *Sox2* transcripts (empty arrowheads). For each gene, the proportion of gastruloids exhibiting the reported pattern is shown. Experimental statistics are shown in Supplementary Information dataset 3. Scale bar, 100 μm. A, anterior; P, posterior; D, dorsal; V, ventral. **d**, Three-dimensional renderings of confocal stacks of 120 h AA gastruloids containing a *Nodal*^{YFP} reporter gene, stained for SOX2 (white) and BRA (red) proteins and imaged from the dorsal (left), ventral (middle) and posterior (right) side of the gastruloid; insets show a confocal section of the posterior region. Reporter-gene expression within the *Bra*-expressing domain on the ventral surface is suggestive of a node-like structure (middle; Extended Data Fig. 6). Additional expression of *Nodal* in a bilaterally asymmetric cluster of cells (white arrowheads) is reminiscent of the asymmetric *Nodal* expression in the embryo. Data are representative of two independent experiments with similar results ($n = 13$). Additional samples are shown in Extended Data Fig. 6e. L, left; R, right. Scale bar, 30 μm. **e**, Bar graph showing the frequency distribution of asymmetric and symmetric expression of *Nodal* or *Meox1* in 120 h AA gastruloids (see Supplementary Information dataset 3 for more details). At this stage, *Nodal* expression was significantly less symmetrical than *Meox1* expression, suggesting that gastruloids may also implement the start of left–right asymmetry. $^{**}P < 0.0001$. The hybrid Wilson/Brown method was used to calculate the confidence intervals.

crest (for example, see ref. [8]) (Fig. 2b, Extended Data Figs. 2b, 3). We also observed an antero-posterior pattern of differentiation along these lineages, reminiscent of the pattern in the embryo. For example, the sequential expression of *Bra*–*Msgn1*–*Meox1*–*Tcf15* recapitulates the spatio-temporal differentiation pattern of paraxial mesoderm[9] (Extended Data Fig. 3a, b). Genes associated with spinal cord development (Extended Data Fig. 3c, d), such as *Irx3*, *Sox1*, *Sox2* and *Lfng* (lunatic fringe) were expressed in a continuous domain along the extension of the gastruloid. Within this domain, *Hes5* and *Dll1* (marking different neural progenitors) were strongly expressed, whereas terminally differentiating cells (*Phox2a*/*Mnx1*-positive) formed an apparent anterior-to-posterior density gradient and were almost completely absent from the posterior-most aspect (Extended Data Fig. 3c, d), reflecting the organization in the embryo[10]. Nevertheless, these ordered patterns of gene expression did not correlate with any precise morphogenesis. For example, neural progenitors did not properly form a neural tube (Fig. 3, Extended Data Figs. 3d, 4a–c), even though sporadic tubular structures were observed (Extended Data Fig. 4a (white arrowheads), c (right)). We observed clumps of cells positive for either SOX1 and OLIG2, SOX1 and PAX3, or SOX1 and PAX7, indicative of dorsal and ventral neural tube progenitors[11] (Extended Data Fig. 4d); however, they lacked a clear segregation along the dorso-ventral extension of the SOX1 domain. Similarly, *Tcf15*-expressing cells did not condense into somites (Extended Data Fig. 3b).

Analysis of different endodermal markers revealed temporal dynamics that were also reminiscent of those in the embryo[12] (Extended Data Figs. 3e, f, 4e, f). Transcripts of *Gsc* and *Cdx2*, markers of definitive endoderm[13,14], were upregulated soon after Chi induction (72 h AA), followed by upregulation of *Cer1* (cerberus) (96–120 h AA) and subsequently *Sorcs2*, *Pax9* or *Shh* (120–144 h AA). All assayed endoderm-expressed genes were active in the ventral-like domain of gastruloids (Extended Data Figs. 3f, 4e, f), resembling expression in the embryo. In a majority of cases, gut-endoderm progenitors appeared as a continuous tubular structure (Extended Data Fig. 4a, e, f; red arrowheads), often spanning the entire antero-posterior extension, reminiscent of an embryonic digestive tract.

We next investigated this unanticipated level of organization and capacity to self-organize an integrated axial system by assessing the expression of genes associated with the developing embryonic axes (Fig. 3). Transcripts of *Wnt3a* and *Cyp26a1* were detected at the caudal extremity of gastruloids, similar to those of *Bra* and *Cdx2* (compare Fig. 3a and Extended Data Fig. 5a with Fig. 1c, d and Extended Data Figs. 3b, 4f), and complementary to the localization of *Raldh2* (also known as *Aldh1a2*) mRNAs (Fig. 3a)—further supporting the existence of an antero-posterior axis. This was also supported by the spatial segregation of the presomitic mesoderm-like domain (marked by *Cyp26a1*) and the *Meox1* somitic mesoderm (Fig. 3a (right), Extended Data Fig. 5a). On the other hand, *Lfng*, *Sox1* and *Sox2* were transcribed in a central and dorsal domain at 144 h AA (Fig. 3b, Extended Data Figs. 3d, 4a), in a complementary fashion to the ventrally located intestinal tract markers (Fig. 3b, Extended Data Figs. 3d, 4e, f, 5b). Additional signs of multi-axial organization were provided by the expression of mesoderm-specific genes *Osr1*, *Pecam1*, *Meox1* and *Pax2* in a medio-lateral symmetry flanking the centrally located *Sox2*-positive domain (Fig. 3c, Extended Data Fig. 3b). Double staining of *Sox2* and *Meox1* (Fig. 3c (right), Extended Data Fig. 5c) and cross-sections (Extended Data Fig. 4b) confirmed the non-overlapping medio-lateral and dorso-ventral distribution of neural and mesodermal progenitors.

Nodal expression was confined to a small and compact region on the ventral most posterior aspect at 120 h AA (Extended Data Figs. 6, 7). These *Nodal*-expressing cells displayed high levels of *Cdh1* (also known as E-cadherin) and dense phalloidin staining (Extended Data Fig. 6a, b), suggestive of a node-like identity[15], a hypothesis supported by the presence of *Nodal* mRNA in a domain comparable to that of *Gsc*, *Bra* and *Chrd* at 96 h AA (Extended Data Fig. 6c, d). *Nodal* mRNA in these cells rapidly decreased and was almost undetectable at 144 h AA. However, despite these signs of a node-like structure, we did not observe any notochord derivatives, which usually originate from the node. The downregulation of *Nodal* in the presumptive node-like cells at 120 h AA coincided with the appearance of patches of *Nodal*-expressing cells along the posterior half of extending gastruloids, which were often distributed in an asymmetric manner (Fig. 3d, Extended Data Figs. 6d, e, 7, Supplementary Videos 3, 4) at 120 h and 144 h AA. *Cer1* also displayed a left–right asymmetric expression, particularly evident at 144 h AA (Extended Data Fig. 6f). This pattern was not observed with *Meox1*, which was predominantly expressed on both sides (Fig. 3d, e, Supplementary Information dataset 3). Together, these data suggest that besides the establishment of a medio-lateral axis, gastruloids may also implement the beginning of left–right asymmetry.

The formation and patterning of post-occipital embryonic territories is associated with the sequential activation of clustered *Hox* genes. Because these genes appeared to be differentially regulated in the RNA-seq time course (Fig. 2a, Extended Data Fig. 2a, b), we assessed whether their sequential activation in time and space[16] was also recapitulated. A pooled PCA analysis using only *Hox* transcripts revealed robust

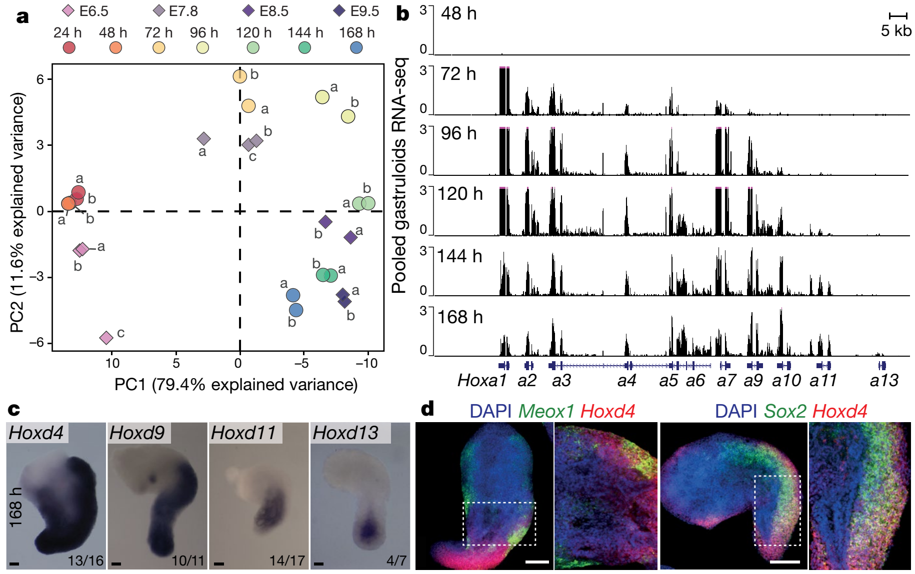


Fig. 4 | Collinear *Hox* gene expression in gastruloids. a, PCA plot based solely on *Hox* transcripts datasets extracted from pooled gastruloids and embryonic data across time points. For gastruloids, $n = 2$ independent biological replicates per time-point; for E6.5 and E7.8 embryos, $n = 3$ independent biological replicates; For E8.5 and E9.5 embryos, $n = 2$ independent biological replicates. **b**, Transcript profiles over the *HoxA* cluster, using time-sequenced pooled gastruloids. A progressive wave of transcription through *Hoxa* genes is observed between the 72 h and 168 h AA time points. The arrangement of the *Hoxa* cluster is shown schematically below the *x* axis. **c**. In situ hybridization of 168 h AA gastruloids using probes for various *Hoxd* genes. Expression becomes spatially restricted along the antero-posterior axis in parallel with the respective position of the genes in the cluster. For each gene, the proportion of gastruloids displaying the reported expression pattern is shown in the bottom right corner of the image, expressed as a fraction of the total number of gastruloids analysed. Scale bar, 100 μm. **d**. Double-FISH staining of *Hoxd4* with *Sox2* or *Meox1* (marking the neural and mesodermal precursors, respectively) showed that *Hoxd4* expression colocalized with both markers, suggesting that gastruloids implement both neural and mesodermal *Hoxd* gene expression. The expression patterns are representative of four independent experiments. Scale bar, 200 μm.

clustering along the time axis (81% variance) and a close correspondence with the dynamic activation of these genes in embryos (Fig. 4a, Extended Data Fig. 8a–c). The variability in *Hox* mRNA content among gastruloids was evaluated using ten individual specimens from three different stages (Extended Data Fig. 9a). Gastruloids at identical time points clustered tightly together based solely on their *Hox* transcripts. Transcript profiles over *Hox* clusters revealed signs of collinear activation, the hallmark of this gene family[17]. In E6.5 embryos, some *Hoxa* and *Hoxd* genes are expressed before gastrulation in extra-embryonic tissues[18] (Extended Data Fig. 8a). Between E7 and E9.5, *Hox* genes start to be transcribed in an order that reflects their 3′ to 5′ position within each cluster (Extended Data Fig. 8a, b). RNA-seq profiling revealed an activation dynamic comparable to that observed in embryos (Fig. 4a, Extended Data Fig. 8c). For instance, whereas *Hoxa* RNA was not detected until 48 h AA, *Hoxa1* to *Hoxa3* expression was robust at 72 h, and was followed by sustained transcription of *Hoxa5*, *Hoxa7* and *Hoxa9* at 96–120 h. *Hoxa10* and *Hoxa11* RNA appeared at 144 h AA, at the same time that *Hoxa1*, *Hoxa2* and *Hoxa3* transcripts started to disappear (Fig. 4b, Extended Data Fig. 8c). Similar dynamics were observed for *Hoxd* genes (Extended Data Fig. 8c–e). The early transcription of 5′ *Hoxa*/*Hoxd* genes (Extended Data Fig. 8a, b) was not observed in gastruloids (Extended Data Figs. 4b, 8c, d), consistent with the absence of extra-embryonic derivatives.

Comparable expression profiles were observed when single organoids were examined (Extended Data Fig. 9a, b), again revealing the reproducibility of this activation process. In the embryo, this temporal activation is paralleled by a collinear distribution of transcripts in space[17]. Likewise, *Hoxa4*/*Hoxd4* displayed an antero-posterior boundary near the anterior aspect of the gastruloid, whereas *Hoxa9*/*Hoxd9*, *Hoxa11*/*Hoxd11* and *Hoxd13* exhibited successively posterior boundaries (Fig. 4c, Extended Data Fig. 9c). Notably, *Hoxd13* transcripts appeared in cells that were located centrally at the posterior extremity, resembling *Hoxd13* expression in the embryonic cloacal area (Fig. 4c). *Hoxa13* expression was also detected at 168 h AA in the posterior aspect, yet rarely (one in 20 gastruloids examined, consistent with the low transcript levels detected in the pooled RNA-seq analysis (Extended Data Fig. 9c). Double staining for *Hoxd4* and either *Sox2* or *Meox1* revealed *Hoxd4* expression in both neural and mesodermal derivatives (Fig. 4d, Extended Data Fig. 5d, e). The implementation in space and time of the *Hox* gene network confirmed the surprisingly high level of organization in the processing of gene regulatory networks, particularly given the absence of extra-embryonic components[19].

We tested the ability of several induced pluripotent stem cell (iPSC) lines to produce gastruloids (Extended Data Fig. 10), observing a similar elongation process in one of them. Thus, iPSCs can generate gastruloids; yet these gastruloids had reduced elongation rates, particularly between 48 and 96 h AA (Extended Data Fig. 10a, b). The expression dynamics of *Bra* were nevertheless similar to their ESC counterparts (Extended Data Fig. 10c, d) and *Cdx2* and the neural markers *Sox1* and *Sox2* were also expressed as in ESC-derived gastruloids (Extended Data Fig. 10d, compare with Fig. 1b, c). Furthermore, iPSC gastruloids implemented temporal and spatial collinear expression of *Hoxd*, albeit with a delay in the onset of expression and a spatial collinearity that was not as clearly organized as in ESC-derived gastruloids (Extended Data Fig. 9d, e).

When compared to single-tissue organoids (for example, see refs.[20–23]), gastruloids exhibit an integrated structure that can specify all major embryonic axes in a coordinated manner, complementing recent reports in which stem cells were used to recapitulate morphological and transcriptional events of early blastocyst, but not of subsequent embryonic stages[19,24]. However, the activation of tissue-specific expression patterns in our gastruloids was not paralleled by a clear internal organization of the corresponding embryonic tissue layers. This observation suggests that caution is needed when considering the possibility of a direct causal relationship between transcriptional programs and early morphogenesis. One potential reason for the low level of tissue organization in gastruloids may be the absence of mechanical interactions and constraints that characterize the developing embryonic context[25]. The autonomy in the patterns of gene expression reported here highlights the potential of gastruloids for studying complex regulatory

circuits, particularly during early post-implantation development and the emergence of body axes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0578-0.

Received: 9 November 2017; Accepted: 13 August 2018;
Published online 3 October 2018.

1. Rossant, J. & Joyner, A. L. Towards a molecular-genetic analysis of mammalian development. *Trends Genet.* **5**, 277–283 (1989).
2. Simunovic, M. & Brivanlou, A. H. Embryoids, organoids and gastruloids: new approaches to understanding embryogenesis. *Development* **144**, 976–985 (2017).
3. van den Brink, S. C. et al. Symmetry breaking, germ layer specification and axial organisation in aggregates of mouse embryonic stem cells. *Development* **141**, 4231–4242 (2014).
4. Sasai, Y., Eiraku, M. & Suga, H. In vitro organogenesis in three dimensions: self-organising stem cells. *Development* **139**, 4111–4121 (2012).
5. Turner, D. A., Baillie-Johnson, P. & Martinez Arias, A. Organoids and the genetically encoded self-assembly of embryonic stem cells. *BioEssays* **38**, 181–191 (2016).
6. Turner, D. A. et al. Anteroposterior polarity and elongation in the absence of extraembryonic tissues and spatially localised signalling in gastruloids, mammalian embryonic organoids. *Development* **144**, 3894–3906 (2017).
7. Zhao, R. et al. Loss of both GATA4 and GATA6 blocks cardiac myocyte differentiation and results in acardia in mice. *Dev. Biol.* **317**, 614–619 (2008).
8. Carninci, P. et al. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
9. Chal, J. et al. Recapitulating early development of mouse musculoskeletal precursors of the paraxial mesoderm in vitro. *Development* **345**, dev157339 (2018).
10. Olivera-Martinez, I. et al. Major transcriptome re-organisation and abrupt changes in signalling, cell cycle and chromatin regulation at neural differentiation *in vivo*. *Development* **141**, 3266–3276 (2014).
11. Dessaud, E., McMahon, A. P. & Briscoe, J. Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network. *Development* **135**, 2489–2503 (2008).
12. Wang, P. et al. A molecular signature for purified definitive endoderm guides differentiation and isolation of endoderm from mouse and human embryonic stem cells. *Stem Cells Dev.* **21**, 2273–2287 (2012).
13. Beck, F., Erler, T., Russell, A. & James, R. Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. *Dev. Dyn.* **204**, 219–227 (1995).
14. Blum, M. et al. Gastrulation in the mouse: the role of the homeobox gene goosecoid. *Cell* **69**, 1097–1106 (1992).
15. Lee, J. D. & Anderson, K. V. Morphogenesis of the node and notochord: The cellular basis for the establishment and maintenance of left–right asymmetry in the mouse. *Dev. Dyn.* **237**, 3464–3476 (2008).
16. Deschamps, J. & van Nes, J. Developmental regulation of the Hox genes during axial morphogenesis in the mouse. *Development* **132**, 2931–2942 (2005).
17. Deschamps, J. & Duboule, D. Embryonic timing, axial stem cells, chromatin dynamics, and the Hox clock. *Genes Dev.* **31**, 1406–1416 (2017).
18. Scotti, M. & Kmita, M. Recruitment of 5′ Hoxa genes in the allantois is essential for proper extra-embryonic function in placental mammals. *Development* **139**, 731–739 (2012).
19. Harrison, S. E., Sozen, B., Christodoulou, N., Kyprianou, C. & Zernicka-Goetz, M. Assembly of embryonic and extra-embryonic stem cells to mimic embryogenesis in vitro. *Science* **356**, eaal1810 (2017).
20. Sato, T. et al. Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. *Nature* **459**, 262–265 (2009).
21. Kadoshima, T. et al. Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell-derived neocortex. *Proc. Natl Acad. Sci. USA* **110**, 20284–20289 (2013).
22. Eiraku, M. et al. Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature* **472**, 51–56 (2011).
23. Lancaster, M. A. et al. Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373–379 (2013).
24. Rivron, N. C. et al. Blastocyst-like structures generated solely from stem cells. *Nature* **557**, 106–111 (2018).
25. Heisenberg, C. P. & Bellaiche, Y. Forces in tissue morphogenesis and patterning. *Cell* **153**, 948–962 (2013).

Acknowledgements We thank members of the Duboule, Lutolf and Martinez-Arias laboratories for sharing material and discussions, J. Deschamps for DNA clones, D. Trono for the *Oct4::GFP* iPSC line, the histology platform and the Gene Expression Core Facility (EPFL) as well as the genomics platform at Geneva University. This work was supported by funds from the BBSRC (No. BB/M023370/1 and BB/P003184/1 to A.M.A.), an NC3Rs David Sainsbury Fellowship (No. NC/P001467/1 to D.A.T.), an Engineering and Physical Sciences Research Council (EPSRC) Studentship (to P.B.-J.), a Company of Biologists Development Travelling Fellowship (DEVTF-151210 to P.B.-J.), a Newnham College Constance Work Junior Research Fellowship (to N.M.), the École Polytechnique Fédérale de Lausanne (D.D. and M.P.L.), the University of Geneva (D.D.), the Swiss National Research Fund (No. 310030B_138662 to D.D.) and the European Research Council grants ERC MOFDH (No. 250316 to A.M.A.), Regul*Hox* (No 588029 to D.D.) and STEMCHIP (No 311422 to M.P.L.).

Reviewer information *Nature* thanks J. Briscoe and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.B., M.G. and D.A.T. produced gastruloids. L.B. performed WISH, imaged and analysed gastruloids, performed RNA extraction and qPCR analysis, analysed and interpreted RNA-seq and wrote the manuscript. N.M. performed bioinformatics analyses, analysed, interpreted and produced figures, and contributed to the manuscript. M.G. contributed to the gastruloid culture protocol, and performed immunostaining and confocal analysis. D.A.T. contributed to the protocols, collected, stained, imaged and analysed gastruloids, and contributed to the manuscript. P.B.-J. contributed to the protocol and collected gastruloids. A.-C.C. cultured gastruloids and performed WISH. M.P.L., D.D. and A.M.A. conceived the project. M.P.L. corrected the manuscript. D.D. and A.M.A. analysed and interpreted the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0578-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0578-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.D. or A.M.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Culture of gastruloids. E14Tg2a (ref. [26]) (E14), *Bra*^{GFP} (ref. [27]), *Gata6*$^{H2B\text{-}Venus}$ (ref. [28]), *Nodal*^{YFP} (ref. [29]) and *Sox1*^{eGFP};*Bra*^{mCherry} double reporter (SBR[30]) mouse embryonic stem (ES) cells were cultured on gelatinized tissue-culture flasks in a humidified incubator (5% CO_2, 37 °C) as previously described[3,6,31]. The E14Tg2A ESC line was obtained from the group of A. Smith. The *Bra*^{GFP} ESC line was obtained from the Keller laboratory. The *Gata6*$^{H2B\text{-}Venus}$ ESC line was generated by the laboratories of A. Martinez Arias and C. Schröter. The *Nodal*^{YFP} cell line was obtained from the Collignon laboratory. The SBR ESC line was obtained through mutual transfer agreement from the laboratory of D. Suter.

The different cell lines used in this study were validated as follows: for the *Bra*^{GFP} ESC line, stimulation with Activin/Chi resulted in an increase of the GFP reporter which overlapped with the signal of the anti-BRA antibody; for the *Nodal*^{YFP} ESC line, we tested whether expression of the reporter was upregulated after Activin stimulation, while its expression was blocked by Nodal inhibitor SB43; the *Gata6*$^{H2B\text{-}Venus}$ and SBR ESC lines were validated by genotyping and via co-staining with GATA6 and BRA antibodies, respectively. All cell lines were routinely tested and confirmed to be free of mycoplasma via the MYCOPLASMACHECK service of GATC Biotech.

The E14, *Bra*^{GFP}, *Gata6*$^{H2B\text{-}Venus}$ and *Nodal*^{YFP} mouse ESC lines were cultured in GMEM supplemented with 10% fetal bovine serum (FBS), non-essential amino acids (NEAA), sodium pyruvate, GlutaMax, beta-mercaptoethanol (β-ME) and LIF (ESL). The SBR cell line was cultured in DMEM supplemented with 10% FBS, NEAA, sodium pyruvate, β-ME, 3 μM Chi[32], 2 μM PD025901 (PD03) and LIF. If cells were not being passaged, half the medium was replaced with fresh medium. Gastruloids were generated as previously described[3,6,33] with modifications (a detailed protocol describing gastruloid culture has been deposited in the Protocol Exchange repository[34]). Mouse ES or iPSCs were collected from tissue-culture flasks, centrifuged and washed twice with warm PBS (containing Ca^{2+} and Mg^{2+}). After the final wash, cells were resuspended in 3–4 ml fresh and warm N2B27 (NDiff) and the cell concentration was determined. For ESCs, the number of cells required to form gastruloids with a diameter of ~150 μm at 48 h AA(~300 cells) was determined and seeded in each well of a round-bottomed, low-adherence 96-well plate as 40 μl droplets of N2B27 (Supplementary Information File 1). For iPSCs, we performed a titration of the initial number of cells required to form gastruloids capable of elongating until at least 144 h AA. Amongst the different conditions tested (200, 400, 600, 800 and 1,200 cells per well) the best results were obtained with a starting number of 800 cells per well. In all cases, a 24 h pulse of 150 μl 3 μM Chi was added at 48 h AA. Medium (150 μl) was replaced with the same volume of fresh N2B27 daily. To extend the culture period, gastruloids were transferred onto low-attachment 24-well plates in 700 μl fresh N2B27 at 120h and cultured in an incubator-compatible shaker for 48 h at 40 r.p.m. Four hundred microlitres medium was replenished at 144 h, and gastruloids fixed at 168 h.

Gastruloids for the different time-points analysed in this study were allocated randomly. Only gastruloids showing clear signs of apoptosis were removed from the experimental group before processing/analysis of the samples. No statistical methods were used to predetermine sample size. Blinding was not relevant to the present study because any specific assumption or hypothesis was postulated a priori, nor was any specific comparison between different treatments or genotypes performed.

Animal experimentation. Mouse embryos for RNA-seq and whole-mount in situ hybridization (WISH) experiments were obtained from CD1 wild-type animals crossed in-house. Adult animals of 3 to 5 months old were used for the crosses. Embryos were collected at E6.5, E7.8, E8.5 and E9.5 dpc. All experiments were performed in agreement with the Swiss law on animal protection (LPA) under license number GE 81/14 (to D.D.) after approval by the Comité Consultatif pour l'Expérimentation Animale du Canton de Genève (the empowered authority).

RNA extraction and RNA-seq libraries preparation and sequencing. Pooled gastruloids derived from either ESCs or iPSCs were collected in 15-ml Falcon tubes and pelleted by centrifuging at 1,000 r.p.m. for 5 min. After washing with cold PBS, gastruloids were resuspended in 100 μl of RNA later (Thermo Fisher) and stored at −80 °C until RNA extraction. The number of ESC gastruloids collected for each RNA-seq replicate varied depending on their size: 295 at 24 h AA; 105 at 48 h AA; 37 at 72 h AA; 14 at 96 h AA and 5 at 120, 144 and 168 h AA. For quantitative PCR analysis, approximately 30 ESC/iPSC-derived gastruloids of 24–96 h, 15 of 120 h and 5 of 144 h AA were pooled for a single RNA extraction. The RNA was extracted using the RNAeasy mini kit (Qiagen), following the manufacturer's instructions. Contaminant DNA was eliminated using on-column DNase digestion (Qiagen RNase free DNase set) and RNA was eluted in RNase-free water. RNA was quantified with the Qubit HS RNA quantification kit (Thermo Fisher) and the integrity was assessed with a Bio-analyser. For each replicate, RNA-seq libraries were prepared from 100 ng pure total RNA using the TruSeq Stranded mRNA protocol from Illumina with poly-A selection. Libraries were sequenced on a HiSeq 2500 machine as single-read, 100 bp reads. For RNA-seq of individual gastruloids, at least 10 from each stage (24 h, 72 h and 120 h AA) were collected individually and transferred to 1.5 ml Eppendorf tubes. After centrifugation, each gastruloid was washed once with cold PBS, resuspended in lysis buffer, and RNA was extracted immediately using the RNAeasy micro kit (Qiagen) with on-column DNase digestion. RNA was quantified with the Qubit HS RNA quantification kit (Thermo fisher) and the integrity was assessed with the Bio-analyser. For each of the replicates of the 24 h AA gastruloids, a total of 0.65 ng RNA was used for the library preparation, whereas 8 ng total RNA was used for the RNA sequencing of the replicates of 72 h and 120 h AA. Libraries were prepared using the SMART-seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech) and the Nextera XT library preparation kit (Illumina), following the manufacturer's instructions, and sequenced on an Illumina NextSeq500 machine as single-read 75-bp reads.

qPCR analysis. Purified RNA from iPSC-derived gastruloids was reverse transcribed using the Promega GoScript Reverse Transcription Kit. Quantitative PCR analysis of mRNA levels for different *Hoxd* genes, *Bra* and the house-keeping gene *Hmbs* was performed using the SYBR Select Master Mix for CFX (Thermo Fisher) kit according to the manufacturer's instructions with specific primers[6,35]. The Biorad CFX96 thermocycler was used. At least two technical (PCR) replicates and two biological replicates were analysed per time point after aggregation.

RNA-seq data processing. No biological replicate was excluded from the RNA-seq analysis. Mitochondrial and non-autosomal genes were excluded as they were not relevant for the biological question addressed in this study. These exclusion criteria were established before the analysis of the data. RNA-seq reads were aligned on the mouse mm10 genome assembly (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.20/) using TopHat 2.0.9[36], implemented in Galaxy[37]. TopHat output files were processed with SAMTools[38] and BedTools[39]. RNA-seq coverages were normalized to the millions of reads mapped for each sample. For the replicates of pooled gastruloids RNA-seq samples, average coverage files were calculated from the normalized coverages of each replicate. We used HTseq[40] implemented in the Galaxy server to count the number of uniquely mapped reads attributable to each gene (based on genomic annotations from Ensembl release 82[41]). We used DESeq2 to perform differential expression analyses. Specifically, we contrasted a generalized linear model that explains the variation in read counts for each gene, as a function of organoid stage, to a null model that assumes no effect of the treatment time. We ran the Wald test and the *P* values were corrected for multiple testing with the Benjamini–Hochberg approach. We computed reads per kilobase of exon per million mapped reads gene-expression levels using Cufflinks[42].

Fragments per kilobase of transcript per million mapped reads (FPKM) levels were $\log_2$-transformed after adding an offset of 1 to each value. The $\log_2$-transformed values were centred across samples before PCA; no variance scaling was performed. For single-gastruloid PCA, only the 1,000 most highly expressed genes were used. For this, an average FPKM expression level of all the replicates of the different time points was calculated and the genes were ordered accordingly.

For cluster analysis of the most variably expressed genes, the top 250 most variable genes were determined by row variance using the genefilter::rowVars function. All heat map clustering, as identified by accompanying dendrogram, was performed using Euclidean distances and complete hierarchical clustering. For each gene cluster, enrichment of gene ontology (GO) terms was performed using Gorilla[43], by comparing the unranked list of gene of each cluster versus the totality of GO-term-annotated genes and by using a *P* value threshold of $P < 10^{-4}$ and a false discovery rate (FDR) <0.05. When more than 10 GO-term categories satisfying these criteria were identified, we used the REVIGO tool[44] to summarize them, using an allowed similarity threshold of 0.7.

Probe cloning, in vitro transcription and in situ hybridization. Specific primers (Supplementary Information File 1) were used to amplify fragments of approximately 400–700 bp of the genes analysed using Toptaq DNA polymerase. The PCR fragments were gel-purified using the Qiagen Gel Extraction Kit and cloned in the pGEM-T Easy Vector System (Promega). Positive clones were verified by standard Sanger sequencing. For antisense RNA probe synthesis, the plasmids were digested with specific enzymes (Supplementary Information File 1) and purified with the Qiagen PCR purification kit. A total of 2 μg of the digested plasmid was used for in vitro transcription using either T7, T3 or Sp6 polymerase (Promega) and the Sigma DIG labelling mix or fluorescein labelling mix (Supplementary Information). The probes were purified using Qiagen RNAeasy mini kit.

Fluorescent and non-fluorescent whole mount in situ hybridization. Gastruloids at different stages AA and E8.5–E9.5 wild-type mouse embryos were collected in 5-ml tubes, fixed overnight at 4 °C in 4% paraformaldehyde (PFA) and stored in methanol at −20 °C. In situ hybridization on whole-mount gastruloids was performed as previously described[45], with some modifications. For non-fluorescent in situ hybridization, gastruloids or embryos were transferred into 12-well Costar Netwell permeable insets (ref. 3477) and rehydrated through a series of decreasing methanol concentrations. After washing in TBST (20mM Tris 137mM NaCl, 2.7mM KCl, 0.1% Tween, pH = 7,4), gastruloids were digested in proteinase K

solution and post-fixed in 4% PFA. The duration and concentration of the proteinase K treatment depended on the developmental stage of the embryo or the gastruloid time point after aggregation. E8 and E9 mouse embryos were incubated for 5 and 7 min in a 5 μg/ml proteinase K solution. Gastruloids at 72–120 h AA and 144–168 h AA were incubated for 1 or 2 min, respectively, in a 1.6 μg/ml proteinase K solution. Proteinase K treatment was stopped by rinsing embryos or gastruloids 3 times in a 2 mg/ml glycine-TBST solution. After post-fixation, gastruloids were prehybridized at 68 °C for 4 h to block non-specific RNA-probe interactions and incubated overnight at 68 °C with specific probes at approximately 200 ng/ml. The next day, probe washes were performed at 68 °C and the gastruloids were transferred to blocking solution at room temperature to impair nonspecific antibody recognition. Subsequently, digoxigenin (DIG)-labelled RNA probes were detected using anti-DIG antibody coupled to alkaline phosphatase (Sigma) at 1:3,000 dilution for 4 h at room temperature. Non-specific antibody background was removed by washing overnight in MABT (100 mM maleic acid, 150 mM NaCl, 0.1% Tween, pH 7.5). The next day, gastruloids were washed 3 times with TBST and 3 times in alkaline phosphatase buffer (0.1 M Tris pH 9.5, 100 mM NaCl, 0.1% Tween) and stained with BM purple solution (Sigma).

For fluorescent in situ hybridization, gastruloids were processed as described above up to the antibody incubation step, but a higher probe concentration was used (500–700 ng/ml). In this protocol, fluoroscein-labelled and DIG-labelled probes targeting each of the two genes to be detected were incubated simultaneously. After probe washing (see above), gastruloids were incubated in blocking solution containing anti-fluorescein antibody coupled to horseradish peroxidase (HRP) 1:100 (Perkin Elmer) for 3–4 h at room temperature. Gastruloids were subsequently washed several times in MABT overnight. The next day, gastruloids were washed 3 times for 5 min in TBST and 3 times in TNT solution (0.1 M Tris pH 7.5, 150 mM NaCl, 0.05 Tween). The fluoroscein-labelled probe was then developed using the TSA PLUS Fluoroscein system (Perkin Elmer) for 10–12 min, following the manufacturer's instructions. To stop the first developing reaction, the anti-fluoroscein-coupled HRP was inactivated by washing the gastruloid 2 times (5 min each) in PBS-Triton 0.3% followed by one hour incubation with PBS-Triton 0.3% + 1% H_2O_2, and post-fixing the gastruloid for 35 min in 4% PFA solution. After 3 washes in TBST and 2 washes in MABT (5 min each), gastruloids were again incubated in blocking solution with anti-DIG antibody coupled to HRP (1:200; Perkin Elmer) for 4 h at room temperature. Subsequently, gastruloids were washed overnight in MABT. The next day, gastruloids were washed 3 times for 5 min in TBST and 3 times in TNT solution. The DIG-labelled probe was then developed using the TSA PLUS Cyanine 3.5 system (Perkin Elmer) for 10 min, following the manufacturer's instructions. The developing reaction was stopped as described above.

Histology. Histology was performed on the EPFL platform (Lausanne). For cryostat sectioning after WISH, gastruloids were placed in Histogel (Thermo Fisher) and oriented under a binocular microscope. Solidified gels were placed in a plastic mould filled with Cryomatrix (Thermo Fisher) and frozen with isopentane. Sections with a thickness of 8 μm were obtained with a Leica CM3050S cryostat. For paraffin section and haematoxylin staining, 5-μm-thick sagittal or coronal sections were used following standard procedures.

Immunostaining and confocal microscopy. Gastruloids were fixed and either Hoechst 33342 or DAPI was used to mark the nuclei. The primary and secondary antibodies used are listed in Supplementary Information File 1. Confocal images of gastruloids were generated using an LSM700 (Zeiss) on a Zeiss Axiovert 200M primarily using a 40× EC Plan-NeoFluar 1.3 NA DIC oil-immersion objective. For some samples, either an EC Plan-NeoFluar 10×/0.3 or a Plan-Apochromat 20×/0.8 was used. Hoechst 33342, Alexa Fluor 488, Alexa Fluor 568 and Alexa Fluor 633 were sequentially excited with 405, 488, 555 and 639-nm lasers, respectively, as previously described[6]. Data capture was carried out using Zen2010 v6 (Carl Zeiss Microscopy) and *z* stacks were acquired with a *z* interval of 0.5 μm. Images were analysed using the ImageJ image processing package FIJI[46]. For FISH samples, FITC and Cy3.5 were excited with 488 and 555 nm lasers, respectively.

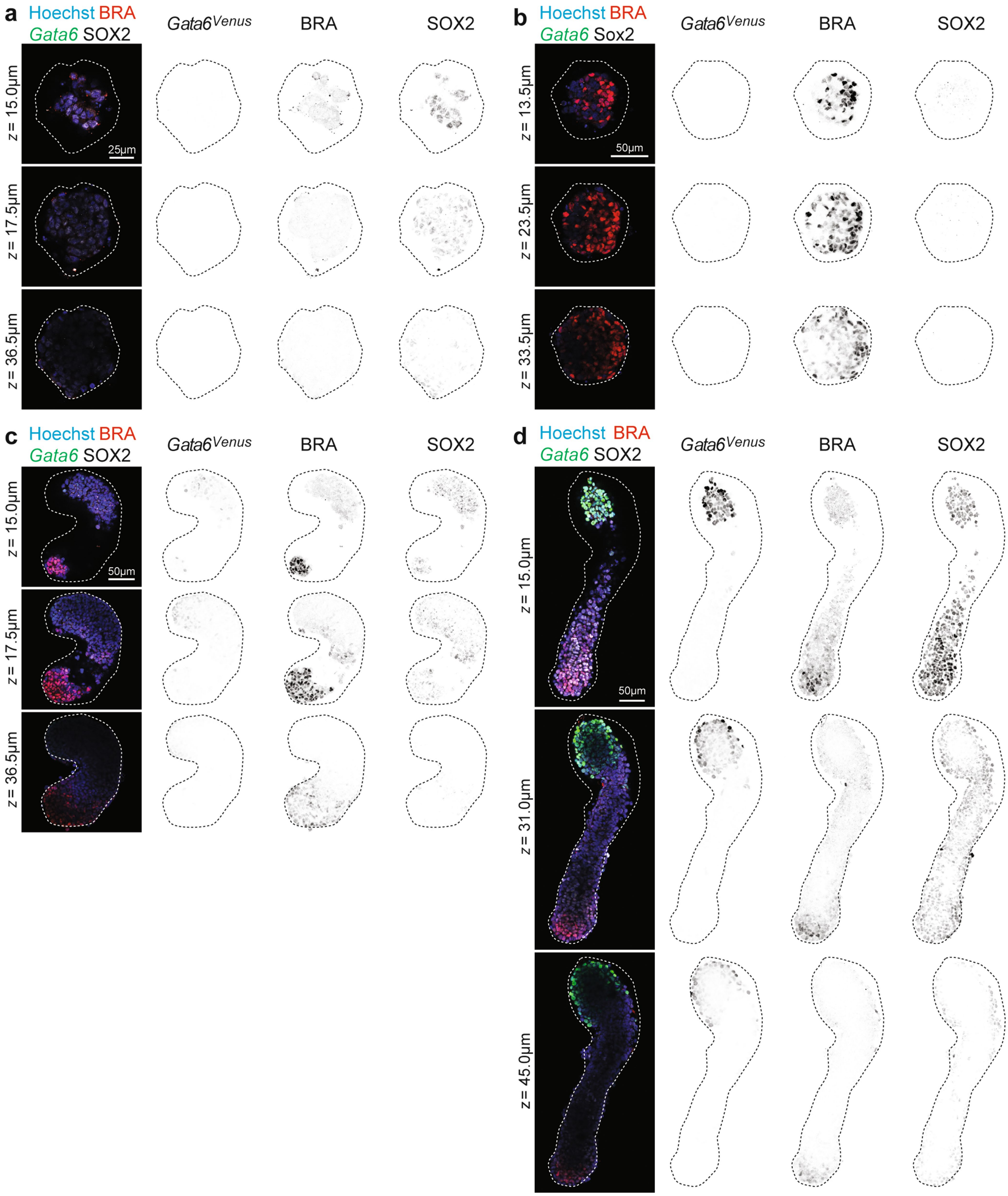
Left–right asymmetry quantification methods. Gastruloids formed from *Nodal*^{YFP} ESCs were fixed at 120 h AA, stained for YFP (indicating nodal expression), *Bra* and *Sox2*, and *z* stacks were acquired from two opposite directions, 0° and 180°. The stacks from both sides of the gastruloid were then aligned and registered in FIJI[46]. Gastruloids were scored as having a node-like structure if a region of nodal expression was found on the ventral surface directly opposite the dorsal expression of *Sox2*, and near the posterior *Bra*-expressing region. The 'left' or 'right' sides of the gastruloid were then inferred from the expression of *Sox2* (dorsal-ventral axis) and *Bra* (antero-posterior axis), and the frequency of those displaying asymmetric *Nodal*^{YFP} expression on the bilateral axis was quantified. To test the significance of this asymmetry, the occurrence of asymmetry in a control gene, *Meox1* (probed by WISH), which is usually expressed on both sides, was quantified and a binomial test of expected versus observed was performed. The fraction of gastruloids displaying symmetric or asymmetric *Nodal* and *Cer1* expression after WISH detection was visually inferred under a stereoscope (Leica MZ205). In these cases, as for *Meox1*, no reference gene was used to determine the left and right side of the gastruloid. In all cases, the frequency of gastruloids displaying symmetric versus asymmetric gene expression was contrasted with the expected frequency based on the expression of these genes in wild-type embryos (*Meox1*: 100% symmetric; *Nodal* and cerberus: 100% asymmetric). The observed proportions of *Nodal* and *Cer1* expression pattern in gastruloids were then compared to those of *Meox1*, using the latter as expected frequency for laterally symmetric expression. The Wilson/Brown hybrid test was used to determine the confidence interval.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

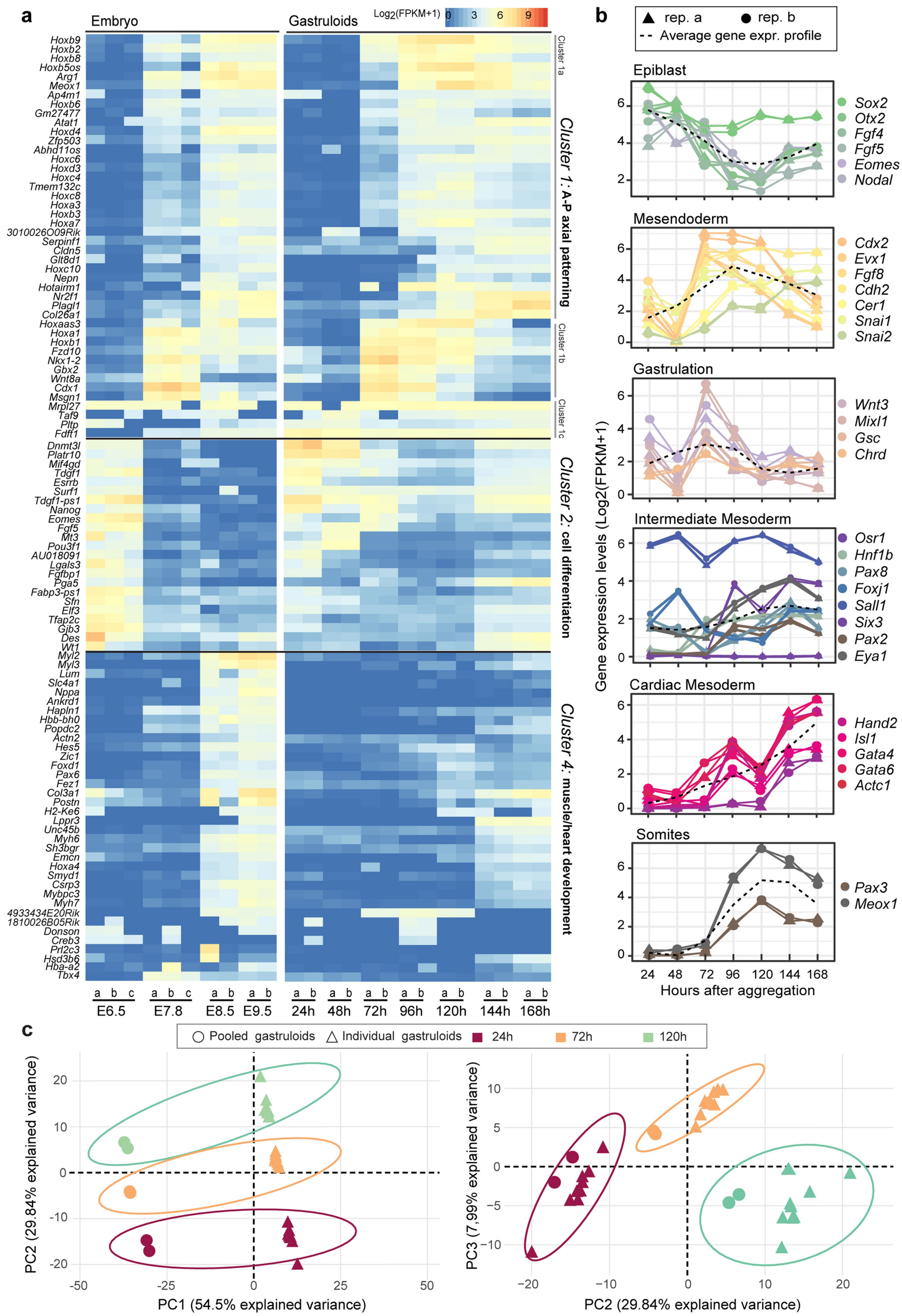
Data availability

All RNA-seq datasets produced in this study are publicly available in the Gene Expression Omnibus (GEO) database under accession code GSE106227 (for gastruloids) and GSE113885 (for embryos). All the scripts used for the analyses of the RNA-seq data are freely available upon request.

26. Hooper, M., Hardy, K., Handyside, A., Hunter, S. & Monk, M. HPRT-deficient (Lesch–Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326**, 292–295 (1987).
27. Fehling, H. J. et al. Tracking mesoderm induction and its specification to the hemangioblast during embryonic stem cell differentiation. *Development* **130**, 4217–4227 (2003).
28. Freyer, L. et al. A loss-of-function and H2B-Venus transcriptional reporter allele for *Gata6* in mice. *BMC Dev. Biol.* **15**, 38 (2015).
29. Papanayotou, C. et al. A novel *Nodal* enhancer dependent on pluripotency factors and Smad2/3 signaling conditions a regulatory switch during epiblast maturation. *PLoS Biol.* **12**, e1001890 (2014).
30. Deluz, C. et al. A role for mitotic bookmarking of SOX2 in pluripotency and differentiation. *Genes Dev.* **30**, 2538–2550 (2016).
31. Turner, D. A. et al. Wnt/β-catenin and FGF signalling direct the specification and maintenance of a neuromesodermal axial progenitor in ensembles of mouse embryonic stem cells. *Development* **141**, 4243–4253 (2014).
32. Ring, D. B. et al. Selective glycogen synthase kinase 3 inhibitors potentiate insulin activation of glucose transport and utilization in vitro and in vivo. *Diabetes* **52**, 588–595 (2003).
33. Baillie-Johnson, P., van den Brink, S. C., Balayo, T., Turner, D. A. & Martinez Arias, A. Generation of aggregates of mouse embryonic stem cells that show symmetry breaking, polarization and emergent collective behaviour *in vitro*. *J. Vis. Exp.* **105**, e53252 (2015).
34. Girgin, M. et al. Generating gastruloids from mouse embryonic stem cells. *Protoc. Exch.* https://doi.org/10.1038/protex.2018.094 (2018).
35. Guerreiro, I. et al. Reorganisation of *Hoxd* regulatory landscapes during the evolution of a snake-like body plan. *eLife* **5**, e16087 (2016).
36. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
37. Blankenberg, D. et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **89**, 19.10.1–19.10.21 (2010).
38. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. Quinlan, A. R. BEDTools: The Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 1–34 (2014).
40. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2014).
41. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016).
42. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).
43. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
44. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
45. Woltering, J. M., Noordermeer, D., Leleu, M. & Duboule, D. Conservation and divergence of regulatory strategies at *Hox* loci and the origin of tetrapod digits. *PLoS Biol.* **12**, e1001773 (2014).
46. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
47. Shawlot, W., Deng, J. M. & Behringer, R. R. Expression of the mouse *cerberus*-related gene, *Cerr1*, suggests a role in anterior neural induction and somitogenesis. *Proc. Natl Acad. Sci. USA* **95**, 6198–6203 (1998).

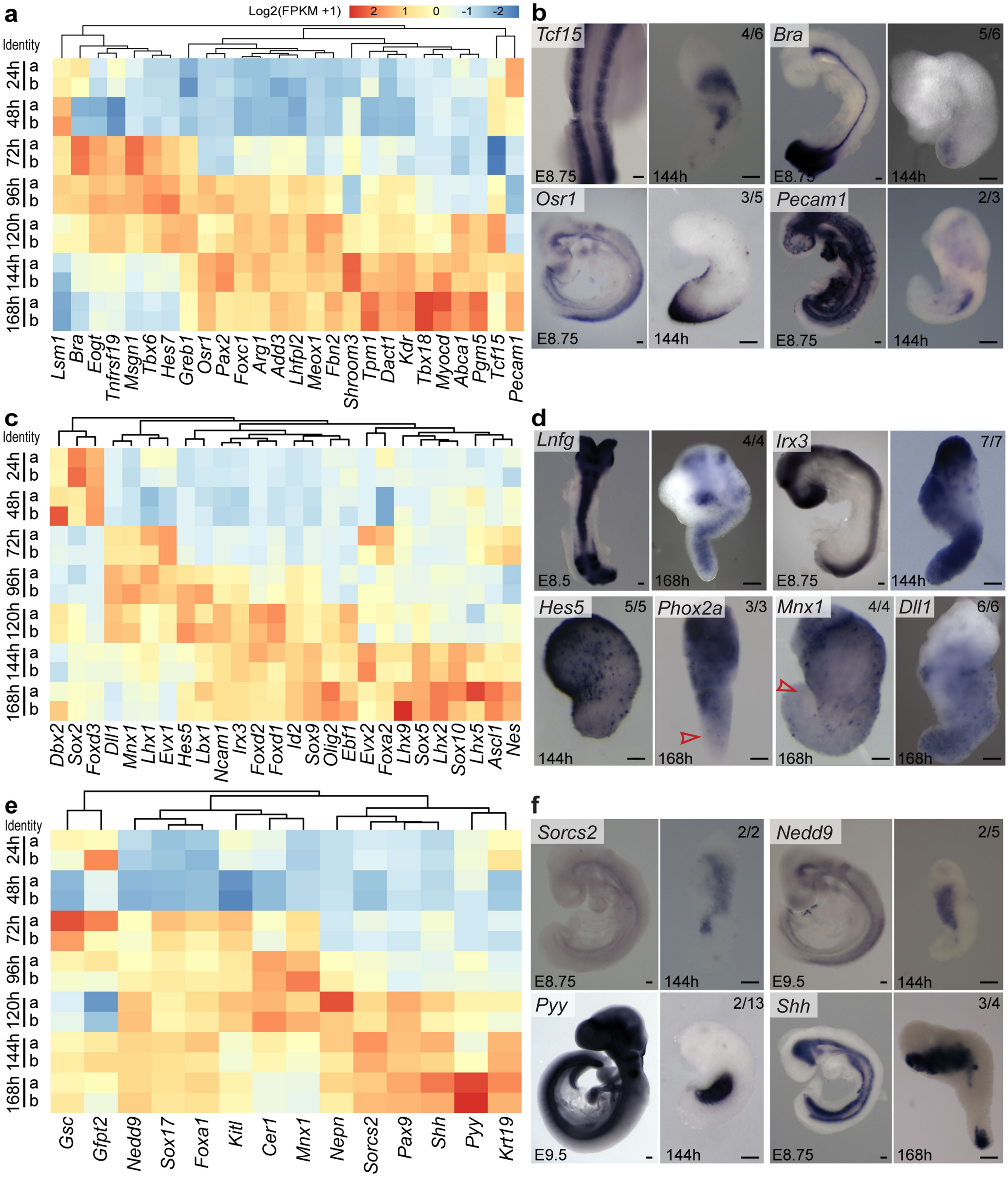


Extended Data Fig. 1 | *z* stacks used for 3D rendering of gastruloids. **a**–**d**, Gastruloids produced using *Gata6*$^{H2B\text{-}Venus}$ ESCs treated with a pulse of the GSK3 inhibitor Chi between 48 h and 72 h AA and fixed at 48 h (**a**), 72 h (**b**), 96 h (**c**) or 120 h AA (**d**) and imaged by confocal microscopy. BRA and SOX2 proteins are stained in red and white, respectively. VENUS signal (green) reports *Gata6* expression and Hoechst 33342 (blue) marks the nuclei. Gastruloids correspond to the 3D renderings shown in Fig. 1a. Each fluorescent channel is displayed to the right of the merged image. *Gata6* (**a**) or *Gata6* and SOX2 (**b**) signals were undetectable, and are therefore not shown. Three *z* sections are shown for each gastruloid. The bright-field outline of each gastruloid is indicated by the dashed lines. Each panel is representative of an experiment perfomed in parallel in seven independent biological replicates showing the same expression pattern. Scale bars are as indicated.



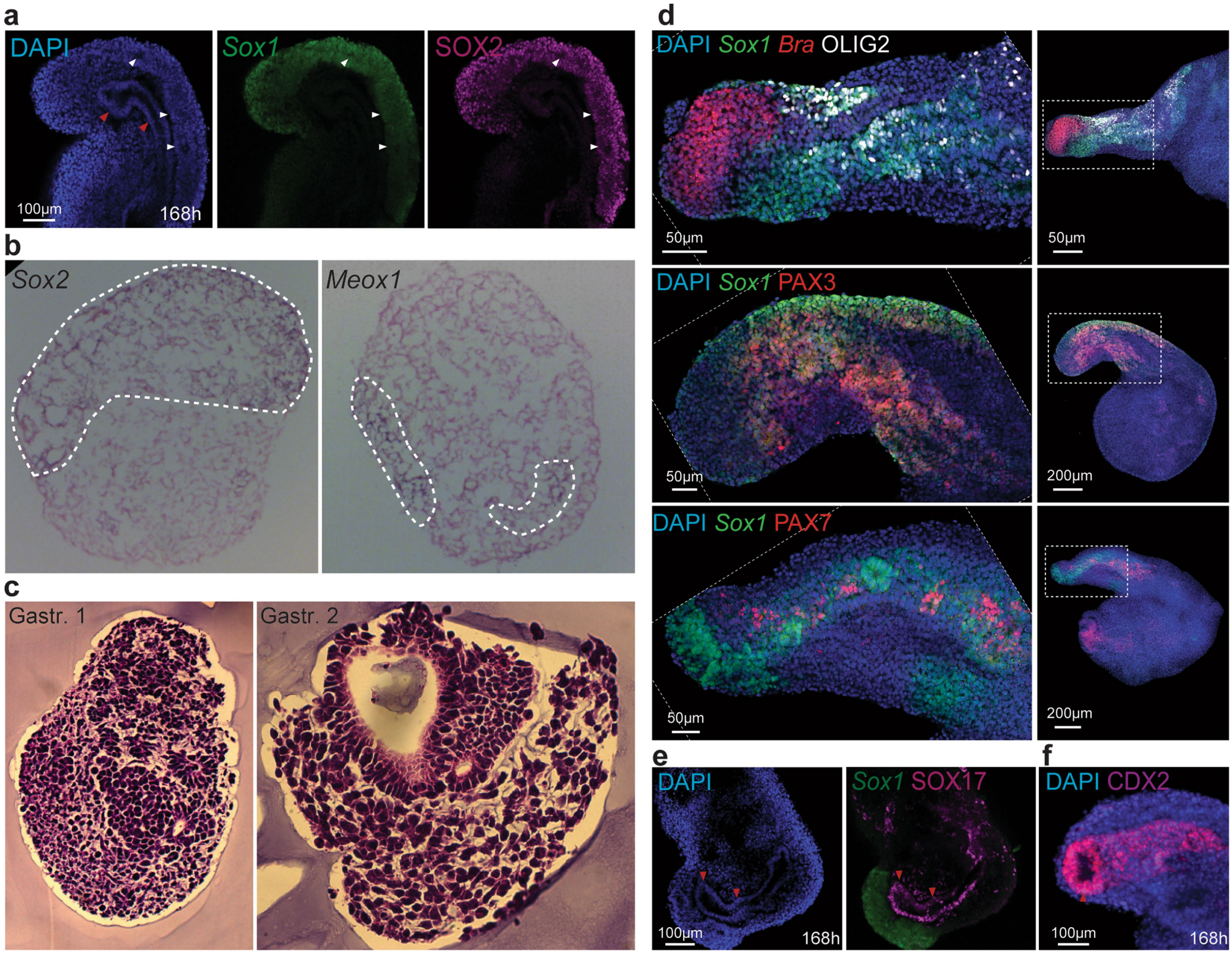
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Transcriptional profiling of mouse embryos and gastruloids. a. Heat map showing the temporal evolution of 97 out of the 250 most variable genes throughout embryonic development from E6.5 to E9.5 (left) and their corresponding expression over the gastruloid time course, from 24 h to 168 h AA (right). Expression levels are indicated by colour scale from blue to red (bottom left). Genes were clustered according to their expression behaviour in the embryo. Enriched GO term categories were identified for each cluster using the Gorilla and REVIGO tools (see Supplementary Information dataset 1). Finally, a functional classification of each cluster was established based on the identified GO term categories and literature-based evidence. **b**, Expression of markers for different embryonic tissues through the gastruloid time course. The two replicates of each time point are represented by a triangle and a circle, respectively. The black dotted line in each plot represents the average behaviour of the genes displayed in the plot. For gastruloids, $n = 2$ independent biological replicates per time point; for E6.5 and E7.8 embryos, $n = 3$ independent biological replicates; for E8.5 and E9.5 embryos, $n = 2$ independent biological replicates. **c**, PCA analysis of RNA-seq datasets from either pooled or individual gastruloids using the top 1,000 most highly expressed genes. Despite different strategies used for RNA-seq of pooled versus individual gastruloids (accounting for the sample segregation across PC1), their clustering illustrates both the homogeneity of gastruloid cultures and the representativeness of pooled samples to single gastruloid samples. For individual gastruloid RNA-seq: $n = 10$ independent biological replicates per time point.

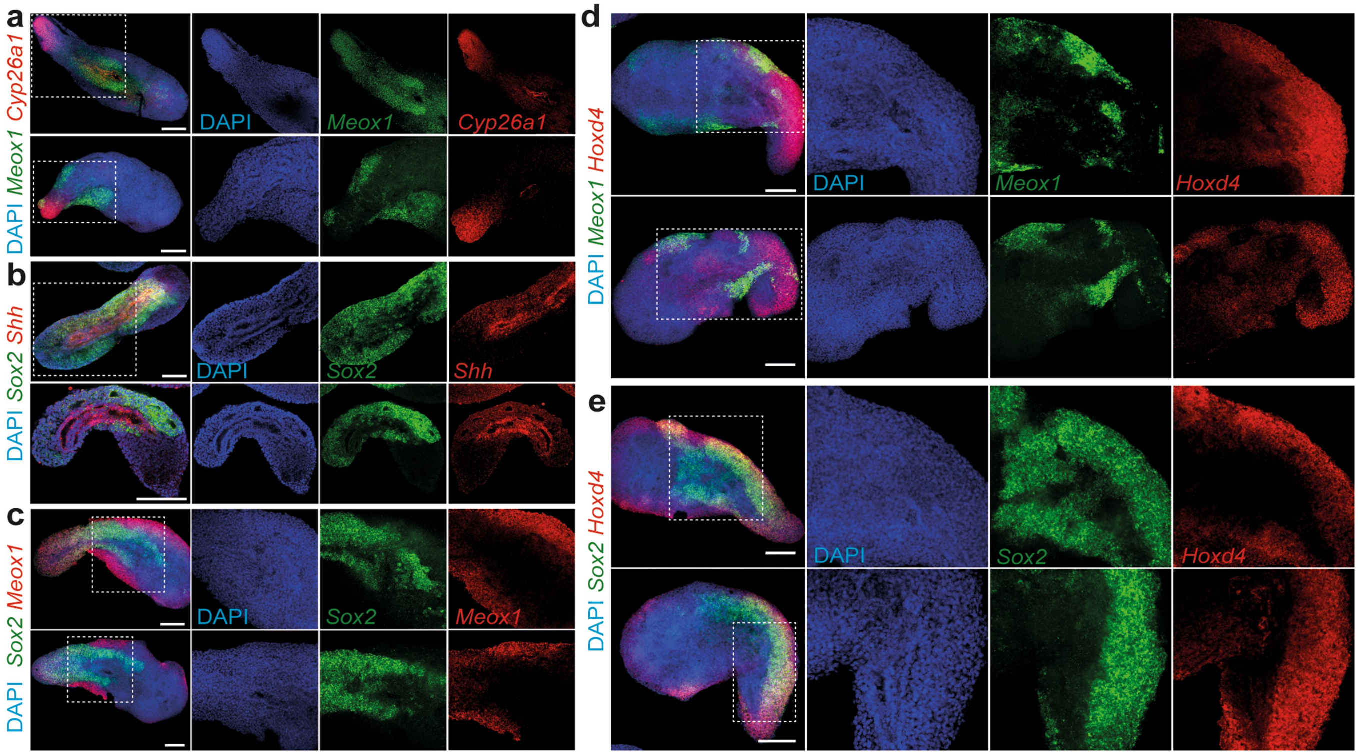


Extended Data Fig. 3 | See next page for caption.

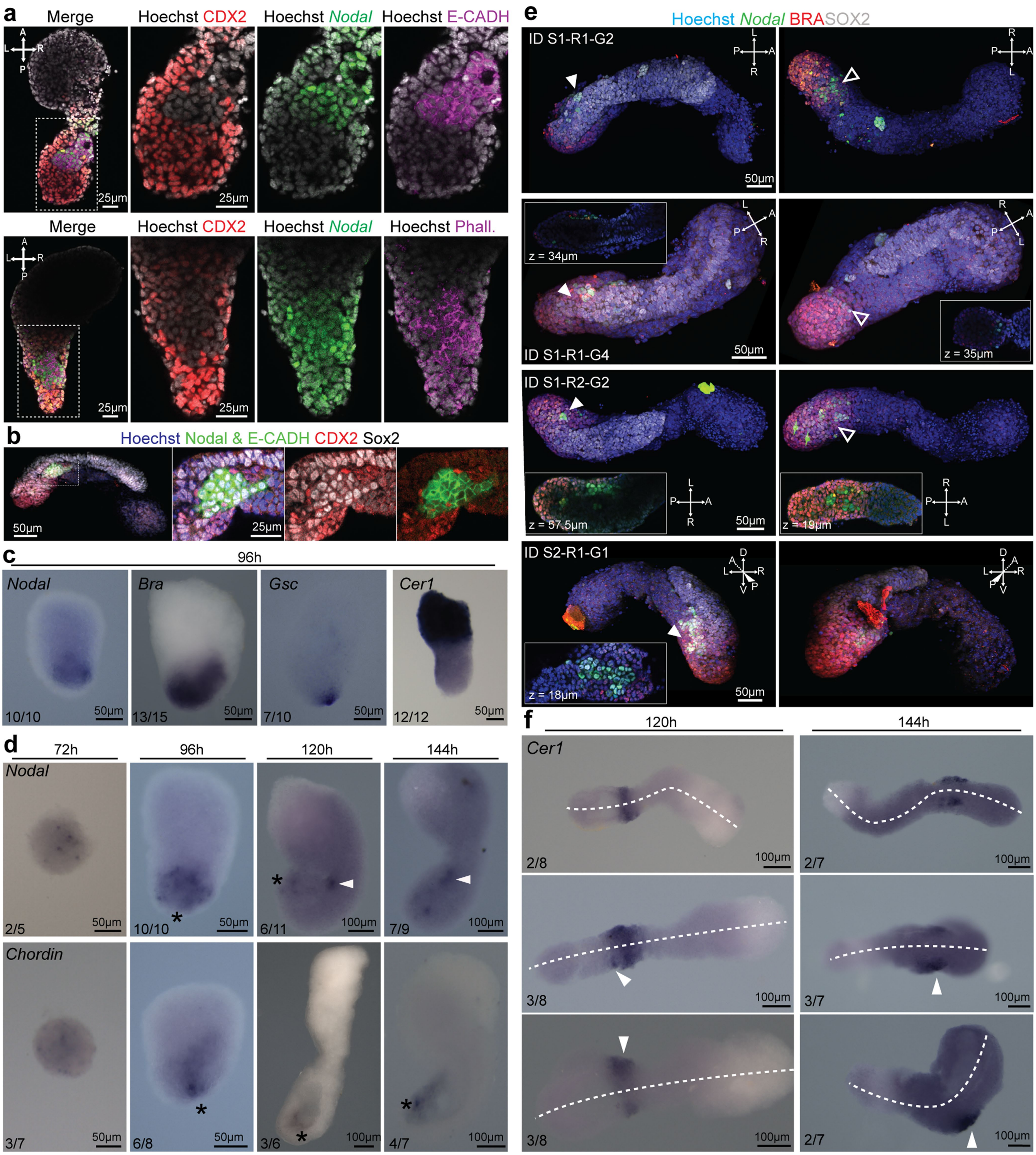
Extended Data Fig. 3 | Gastruloids display spatio-temporal organization in the expression profiles of neural, mesodermal and endodermal marker genes. **a**–**f**, The expression profiles of several genes usually expressed in the embryonic neural, mesodermal and endodermal domains were analysed by plotting RNA-seq data from the pooled gastruloids in heat maps of scaled gene expression (2 independent biological replicates per time point) (**a**, **c**, **e**) and/or by WISH (**b**, **d**, **f**). **a**, **b**, Genes usually expressed in different types of mesoderm precursors in the embryo (for example, *Tcf15* in paraxial somatic mesoderm, *Osr1* in intermediate mesoderm, *Bra* in tail bud, notochord and presomitic mesoderm, and *Pecam1* in lateral plate mesoderm) were expressed in reproducible and spatially restricted domains within the gastruloids. **c**, **d**, Expression of different neural markers was detected in our RNA-seq (**c**). Transcripts of genes such as *Lfng* or *Irx3* formed continuous and homogenous domains located in the central and dorsal portion of the gastruloids, reminiscent of their expression domains in the embryo (**d**, top panels). Genes involved in Notch signalling in neural progenitors (*Hes5*, *Dll1*) and in the terminal differentiation of neural precursors (*Phox2a*, *Mnx1*) displayed a salt-and-pepper expression pattern, consistent with the lack of an organized neural-tube structure (see also Extended Data Figs. 4a, c, 5). However, the latter mRNAs also displayed a graded distribution along the anterior-to-posterior extension of the gastruloid axis and were absent from its posterior half (empty red arrowheads). **e**, **f**, Endoderm-specific genes were also expressed in gastruloids. In particular, genes expressed in the embryonic digestive tract were consistently found on the ventral side of gastruloids. For each gene, the proportion of gastruloids displaying the reported expression pattern is shown in the upper right corner of the image, expressed as a fraction of the total number. Experimental statistics are provided in Supplementary Information dataset 3. Scale bar, 100 μm.



Extended Data Fig. 4 | Tissue organization in gastruloids. **a**, Gastruloids formed from *Sox1*^{GFP};*Bra*^{mCherry} (SBR) line and stained for *Sox2* expression (*Sox1*^{GFP} and SOX2 signals are displayed in green and magenta, respectively). White arrowheads indicate tubular SOX2/*Sox1*-positive neural structures. Red arrowheads point to the presumptive digestive tube. **b**, WISH on 8-μm transverse cryosections of gastruloids at 144 h AA using *Sox2* and *Meox1* antisense probes, counter-stained with Nuclear Fast Red. *Sox2*-positive cells localized predominantly in a compact dorsal domain, whereas *Meox1* signals were found in two bilateral domains. The domain of expression of each gene is outlined with white dashed lines. **c**, Haematoxylin and eosin staining of transverse paraffin sections of different gastruloids at 120 h AA, showing the diversity of cell types and several levels of tissue organization. **d**, Gastruloids formed from *Sox1*^{GFP};*Bra*^{mCherry} ESCs were fixed and stained at 168 h AA for OLIG2 (top, white), PAX3 (middle, red) and PAX7 (bottom, red). Scale bars as indicated. **c**, **d**, Gastruloids formed from *Sox1*^{GFP}; *Bra*^{mCherry} ESCs collected at 168 h AA and stained for SOX17 (magenta, **c**) or CDX2 (magenta, **d**). Scale bars as indicated. All immunostaining experiments were repeated twice, with three biological replicates per experiment, with similar results.

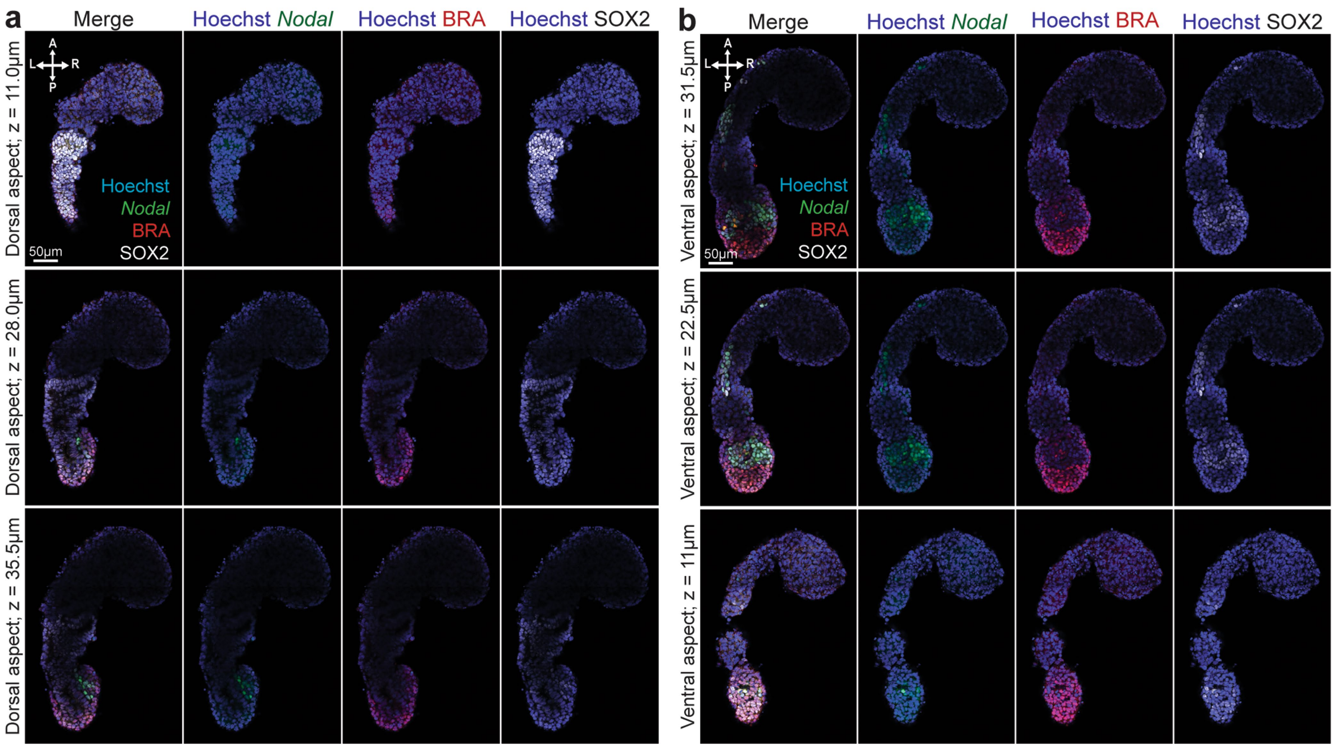


Extended Data Fig. 5 | Double-FISH staining shows organized gene expression across the three main gastruloid axes. a–e, Double-FISH staining of gastruloids at 144 h AA with *Meox1* and *Cyp26a1* (**a**), *Sox2* and *Shh* (**b**), *Sox2* and *Meox1* (**c**), *Meox1* and *Hoxd4* (**d**) or *Sox2* and *Hoxd4* (**e**). **a–e**, Experiments were repeated twice in three biological replicates with similar results. Scale bar, 200 μm.

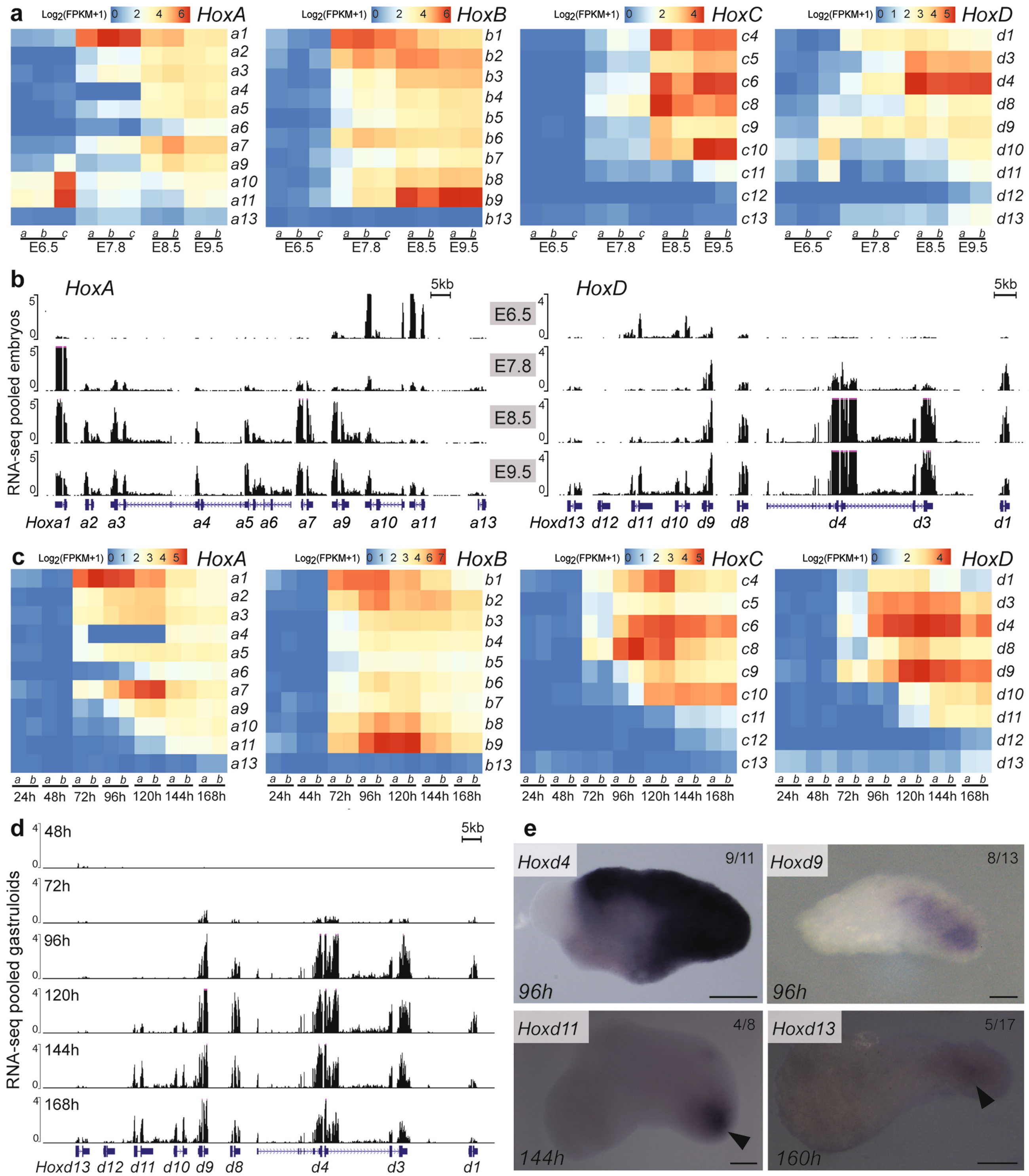


Extended Data Fig. 6 | See next page for caption.

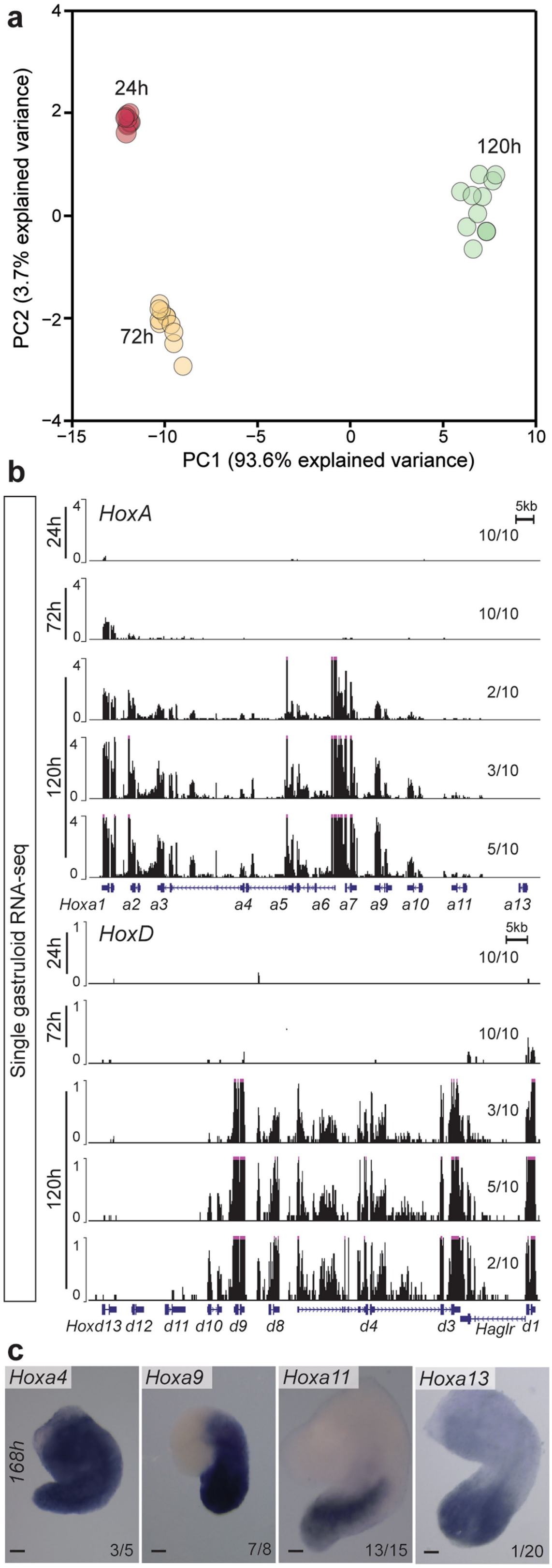
Extended Data Fig. 6 | A node-like structure and left–right asymmetry in gastruloids. **a**, **b**, Gastruloids formed from *Nodal*^{YFP} ESCs were fixed at 120 h AA. They were stained for CDX2, YFP (*Nodal*^{YFP}) and E-cadherin (**a**, top panel), CDX2, YFP (*Nodal*^{YFP}, green) and phalloidin (**a**, bottom panel) or for CDX2, YFP and E-cadherin (both with an Alexa-488 secondary antibody), and SOX2 (**b**). Maximum intensity projection of a representative gastruloid in **b**, with the node-like structure highlighted. Hoechst 33342 labels the nuclei (greyscale in **a**, blue in **b**). Data are representative of one experiment performed in three independent biological replicates. **c**, **d**, In situ hybridization showing expression of the indicated genes in gastruloids at different time-points AA. *, presumptive node-like cells. White arrowheads point towards *Nodal*-expressing cells distributed asymmetrically, on the lateral side of the gastruloid. Whereas *Nodal* was expressed in the presumptive node region from 96 h AA, no clear asymmetry in transcript distribution was observed at that stage. **e**, Three-dimensional renderings of confocal stacks of 120 h gastruloids containing a *Nodal*^{YFP} reporter gene (green) and stained for SOX2 (white) and BRA (red) proteins. SOX2 signal identifies dorsal cells. Left and right panels show the same gastruloid, imaged from two different polar directions that is, top (dorsal) and bottom (ventral) or 'left' and 'right', depending on the orientation of the gastruloid. Insets in specific panels show a cross-section through the gastruloid at the indicated *z* plane. White arrowheads indicate the region of biased *Nodal* expression. Empty white arrowheads point to the node-like cells marked by the *Nodal*^{YFP} reporter gene (see also Fig. 4d). These results are consistent with the asymmetric distribution of *Nodal* transcripts at 120–144 h AA. **f**, In situ hybridization showing expression of *Cer1* in 120 h AA (left) and 144 h AA (right) gastruloids. The gastruloid midline is marked by a dashed white line. At this stage, *Cer1* is expressed in the presumptive embryonic somitic territory[47] and the pattern in gastruloids may reflect this specificity. In **c**, **d** and **f**, the proportion of gastruloids displaying the reported expression pattern is shown at the bottom left corner of each image, expressed as a fraction of the total number of specimens analysed (see Supplementary Information dataset 3 for a complete statistical report).



Extended Data Fig. 7 | *z* stacks used for 3D rendering of gastruloids. **a**, **b**, Dorsal (**a**) and ventral (**b**) sections of the same representative gastruloid shown in the 3D renderings in Fig. 3d, fixed and stained at 120 h for *Nodal*^{YFP} (green), BRA (red) and SOX2 (white). Hoechst 33342 was used to label nuclei. Data are representative of two independent experiments with $n = 13$ biological replicates in total (see Supplementary Information dataset 3 for a detailed statistical report). Scale bar, 100 µm.

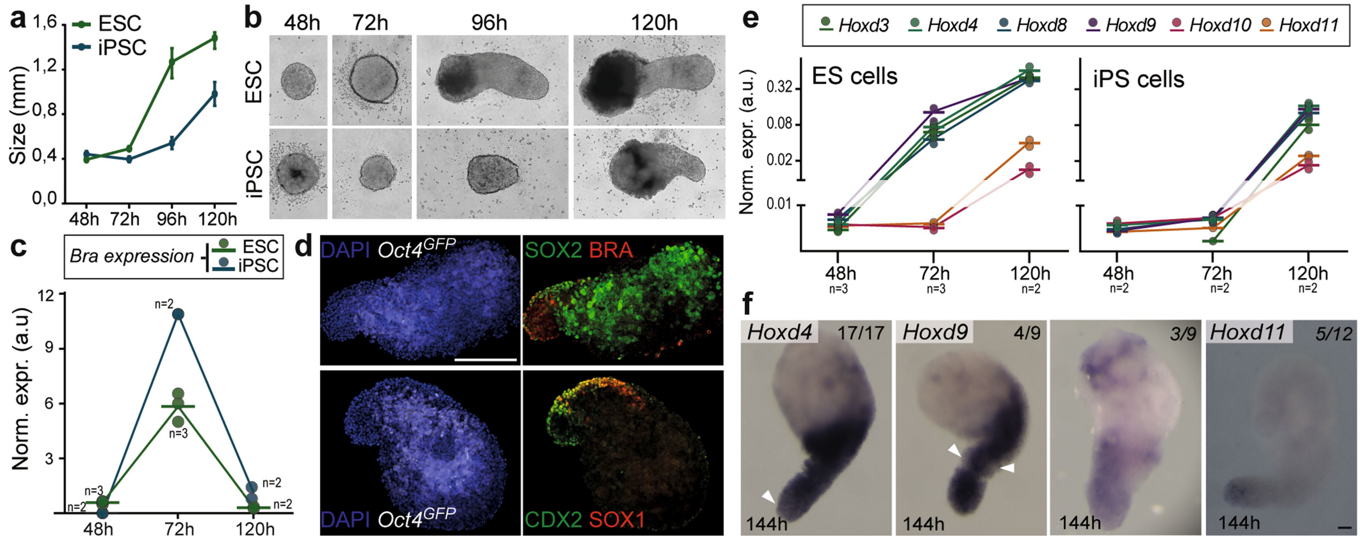


Extended Data Fig. 8 | *Hox* expression profiles in mouse embryos and gastruloids. a, Heat map of unscaled gene expression in E6.5–E9.5 mouse embryos, showing levels of *Hox* gene transcripts over time. Between 2 and 3 independent biological replicates were used for each time point (indicated below each graph). **b**, RNA-seq mapping showing *Hoxa* and *Hoxd* gene expression in these embryos. After a first wave of transcription of 5′ *Hoxa* and *Hoxd* genes, which is likely to reflect their activation in extra-embryonic tissues, the *HoxA* and *HoxD* clusters were progressively transcribed between E7.8 and E9.5, when expression of *Hox13* paralogues was detected. Each profile was averaged from independent biological replicates indicated in **a**. **c**, Heat map of unscaled gene expression in pooled gastruloids, showing *Hox* gene transcript levels over time. Two independent biological replicates were used per time point. **d**, RNA-seq mapping showing *Hoxd* gene expression in pooled gastruloids at different time points. Sub-groups of *Hoxd* genes are progressively activated between 72 h and 168 h AA, when expression of *Hoxd13* starts to be detected (**e**). This resembles the temporal activation described in vivo (**a**, **b**). Each profile represents the average of two independent biological replicates. **e**, WISH of gastruloids collected at different time points, showing the detectable initiation of expression of different *Hoxd* genes. Each panel shows the earliest stage at which the indicated gene was detected (black arrowhead). Expression of *Hoxd4* was already strong at 96 h AA, indicating that its transcripts are rapidly upregulated compared to *Hoxd9*, which is expressed at low levels at this stage. Scale bar, 100 µm. The fraction of gastruloids displaying the reported expression pattern is indicated in the upper right corner of each image. Experimental statistics are provided in Supplementary Information dataset 3.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Homogeneity in *Hox* transcript profiles for individual gastruloids. a, PCA based on *Hox*-transcript datasets only, extracted from individually sequenced gastruloids across time points (10 individual organoids per time point, representing independent biological replicates). The analysis was carried out using the $\log_2$-transformed FPKM+1 value of all 39 *Hox* genes. Replicate batches of organoids primarily cluster according to their age at collection. The clustering revealed the low sample-to-sample variation. however, replicates were clearly separated by the temporal parameter, representing 93.6% of total sample variation. **b**, Comparison of *Hoxa* (top) and *Hoxd* (bottom) gene-expression profiles among individual gastruloids confirmed the low inter-sample variation among time points, illustrated with the 120 h AA condition. **c**, WISH of 168 h AA gastruloids showing the expression of different *Hoxa* paralogues. The proportion of gastruloids displaying the reported expression pattern is shown in the upper right corner of the image, expressed as a fraction of the total number. Experimental statistics are provided in Supplementary Information dataset 3. Scale bar, 100 μm.



Extended Data Fig. 10 | Axial extension and spatio-temporal Hox expression patterns in iPSC-derived gastruloids. a. Dot plot representing the progression in the measured longitudinal extension of gastruloids produced either from ESCs or from iPSCs. In each case, 10 different gastruloids were measured at the different time points indicated. The median (round points) and the interquartile range (vertical bars) are reported. **b**, Light microscopy images showing representative examples of gastruloids at the different time points analysed in **a**. Zoom: 10×. Note that iPSC-derived gastruloids exhibit delay in their longitudinal extension rate and at 120 h AA they are markedly smaller than their ESC-derived counterparts. For this analysis, gastruloids were produced starting from the same number of cells (800 cells per well). **c**, Dot plots representing the *Bra* mRNA levels, showing comparable dynamics of this gene in both types of gastruloids. Circles represent individual data points and the short horizontal line represents the mean. The number of biological independent replicates (*n*) per condition is indicated. **d**, Confocal images showing the expression of *Oct4*, SOX2 and BRA (top) or of *Oct4*, SOX1 and CDX2 (bottom) in 120 h AA gastruloids derived from the iPSC line *Oct4::Gfp* (IpSL40N). iPSC-derived gastruloids were fixed and stained for SOX2 and BRA (top) and CDX2 and SOX1 (bottom). *Oct4::GFP* signal is shown in grey. Scale bar, 200 μm. In each case, data are representative of one experiment with three independent biological replicates. **e**, Dot plots representing the *Hoxd* mRNA levels in ESC- or iPSC-derived gastruloids collected at different time points AA. Each circle represents an independent biological replicate, the horizontal bars represent the mean value of the replicates. Both sets of gastruloids sequentially activated *Hoxd* gene expression. However, their temporal activation seemed to be delayed in iPSC gastruloids (especially that of the most 3′ *Hoxd* paralogues). **f**, WISH of 144 h AA gastruloids showing the expression of different *Hoxd* paralogues. Even though iPSC-derived gastruloids reproduced the antero-posterior *Hoxd* collinear expression, the *Hoxd9* expression domain often extended more anteriorly in comparison to that in ESC-derived gastruloids (see Fig. 4c), occupying roughly the same domain as *Hoxd4*. Patches of *Hoxd*-negative cells were often observed within the *Hoxd4*/*Hoxd9* expression domain (white). The fraction of gastruloids displaying the reported expression pattern is indicated in the upper right corner of each image. Experimental statistics are provided in Supplementary Information dataset 3. Scale bar, 100 μm.

https://doi.org/10.1038/s41586-018-0557-5

Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate

Januka S. Athukoralage[1,2], Christophe Rouillon[1,2], Shirley Graham[1], Sabine Grüschow[1] & Malcolm F. White[1]*

The CRISPR system provides adaptive immunity against mobile genetic elements in prokaryotes, using small CRISPR RNAs that direct effector complexes to degrade invading nucleic acids[1–3]. Type III effector complexes were recently demonstrated to synthesize a novel second messenger, cyclic oligoadenylate, on binding target RNA[4,5]. Cyclic oligoadenylate, in turn, binds to and activates ribonucleases and other factors—via a CRISPR-associated Rossman-fold domain—and thereby induces in the cell an antiviral state that is important for immunity. The mechanism of the 'off-switch' that resets the system is not understood. Here we identify the nuclease that degrades these cyclic oligoadenylate ring molecules. This 'ring nuclease' is itself a protein of the CRISPR-associated Rossman-fold family, and has a metal-independent mechanism that cleaves cyclic tetraadenylate rings to generate linear diadenylate species and switches off the antiviral state. The identification of ring nucleases adds an important insight to the CRISPR system.

Cyclic oligoadenylate (cOA; with a ring size of 4 (tetraadenylate, cA_4) or 6 (hexaadenylate, cA_6) AMP monomers) has emerged as a key second messenger that signals the presence of invading mobile genetic elements in prokaryotes that contain type III (Csm or Cmr) CRISPR systems[4,5]. cOA is synthesized by the cyclase domain of the Cas10 subunit, which is activated by the binding of target RNA. cOA, in turn, activates a range of proteins with CRISPR-associated Rossman-fold (CARF) domains including 'higher eukaryotes and prokaryotes nucleotide binding' (HEPN)-domain ribonucleases (Csm6 and Csx1) and transcription factors[6–8], which precipitates an antiviral state in infected cells that may enhance viral clearance, result in dormancy or result in cell death[9]. There are interesting parallels with the cGAS–cGAMP–STING pathway in eukaryotes, in which detection of DNA in the cytoplasm leads to synthesis of a cyclic nucleotide that activates the host immune response[10]. The HEPN ribonucleases are important for CRISPR-based immunity in vivo[11–13]. cOA is therefore a potent signalling molecule that must be tightly controlled if cells are to survive a viral infection. cOA synthesis is switched off when type III effectors cleave and release viral RNA targets[5,14]. However, this will not remove extant cOA, which could potentially lead to untrammelled RNase activity and cell death. Other cyclic nucleotide signalling molecules in prokaryotes, such as di-cAMP, are degraded by specific phosphodiesterases[15,16], but the enzyme specific for cOA is unknown. Here we report the identification and characterization of a family of enzymes that degrade cA_4, providing a mechanism for the deactivation of the antiviral state induced by cA_4 in cells that contain a type III CRISPR system.

We undertook a classical biochemical approach to identify the enzyme responsible for the degradation of cA_4. We started with a cell lysate from the crenarchaeote *Sulfolobus solfataricus* (abbreviated Sso in uncharacterized proteins), and noted the presence of an activity that converted radioactively labelled cA_4 into a form (hereafter referred to as 'product X') that migrated more slowly in denaturing gel electrophoresis (Fig. 1). We fractionated the cell lysate through three chromatography steps (phenyl-sepharose, size exclusion and heparin) and followed the activity. The final purification step was followed by assay of each fraction and then by sodium-dodecyl-sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) analysis (Fig. 1d), which revealed that the activity correlated with a single protein that was identified by mass spectrometry as Sso2081—a member of the CARF-domain-containing

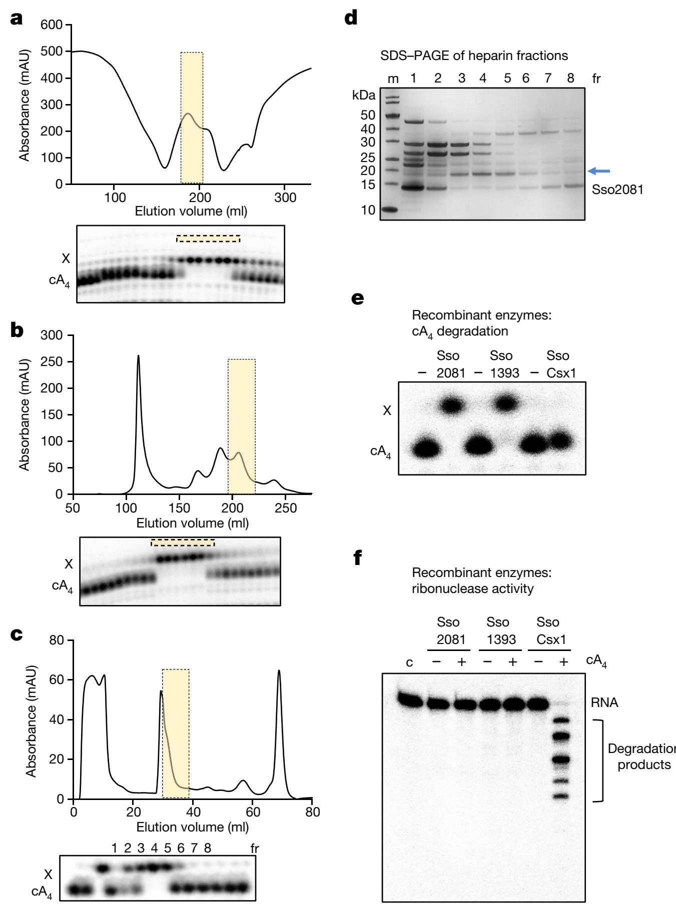


Fig. 1 | Purification and identification of the enzyme that degrades cA_4. a–c, *S. solfataricus* cell lysate was fractionated by phenyl-sepharose (**a**), size-exclusion (**b**) and heparin (**c**) chromatography. At each stage, fractions were assayed for cA_4 conversion activity using radioactive cA_4, and active fractions that generated product X (indicated by shaded boxes) were pooled for the next stage. **d**, Following the heparin column, each fraction was assayed and analysed by SDS–PAGE. The band corresponding to the peak of activity (arrowed) was excised from the gel and identified by mass spectrometry as Sso2081, a CARF-domain protein. **e**, Purified recombinant Sso2081 and Sso 1393 degrade cA_4, but Csx1 does not. **f**, Only Csx1 degrades linear RNA in the presence of cA_4. Panels **a**–**f** are representative of experiments performed at least in duplicate. c, control without protein; mAU, milli-absorbance units; m, protein markers; fr, fractions.

[1]Biomedical Sciences Research Complex, School of Biology, University of St Andrews, St Andrews, UK. [2]These authors contributed equally: Januka S. Athukoralage, Christophe Rouillon. *e-mail: mfw2@st-andrews.ac.uk

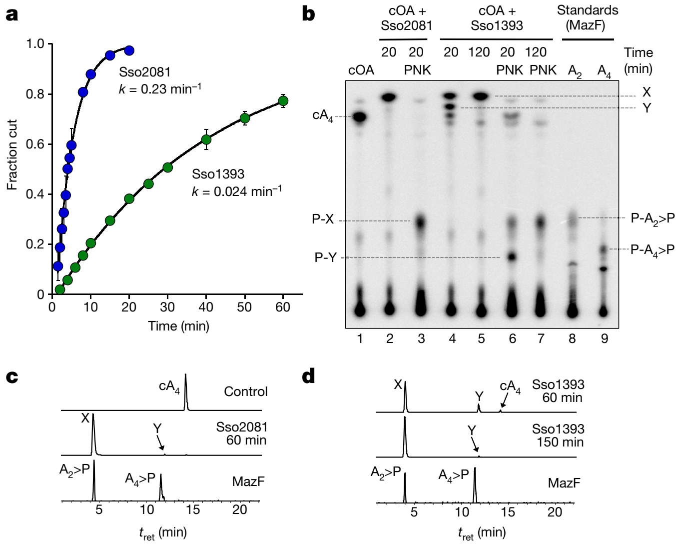


Fig. 2 | Kinetics and products of ring nuclease activity. a, Single-turnover kinetic analysis of cA_4 cleavage by Sso2081 and Sso1393. Sso2081 is a tenfold-faster enzyme. Data points are the means of triplicate measurements, with standard deviation shown, and are technical replicates representative of duplicate experiments. **b**, Thin-layer chromatography analysis of substrates and products. Lane 1 shows cA_4 synthesized by the Csm complex. Lane 2 shows the product X of conversion by Sso2081 after 20 min. Lane 3 shows product X after phosphorylation (P-X) by polynucleotide kinase. Lanes 4–7 show the products X and Y of cA_4 conversion by Sso1393 after 20 and 120 min, respectively, before and after polynucleotide kinase treatment (P-Y, phosphorylated product Y). Lanes 8 and 9 are markers for P-A_2>P and P-A_4>P, generated by the MazF nuclease. This gel is representative of duplicate experiments. **c**, **d**, LC–HRMS analysis of cA_4 cleavage by Sso2081 and Sso1393, respectively, showing ion chromatograms extracted for m/z 657.10 ± 0.5, corresponding to cA_4^{-2}, A_4>P^{-2} and A_2>P^{-1}. The control is cA_4 incubated for 150 min without enzyme. Traces for the conversion of cA_4 by Sso2081 after 60 min, and by Sso1393 after 60 and 150 min are shown. Products X and Y are indicated. The bottom traces show the linear oligoadenylates A_2>P and A_4>P, generated by MazF, as standards. The data are representative of experiments performed in at least duplicate. PNK, polynucleotide kinase; t_{ret}, retention time.

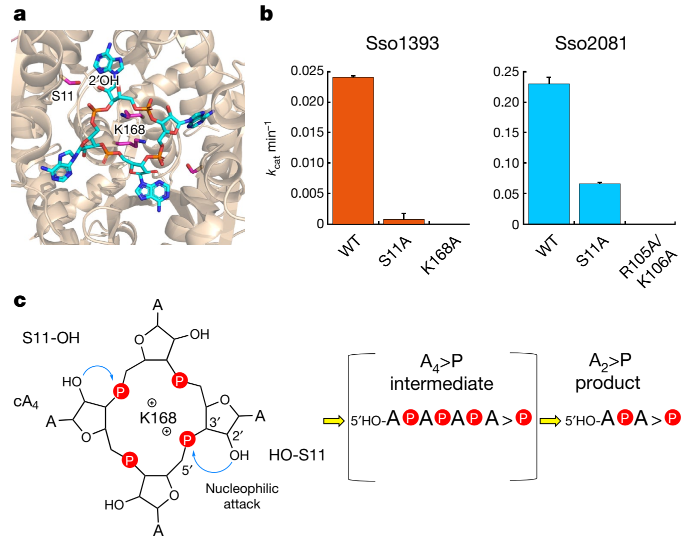


Fig. 3 | Structure and mechanism of ring nucleases. a, Structure of the CARF domain of Sso1393 with cA_4 docked. The active site residues K168 and S11 are shown. **b**, Kinetic analysis of Sso1393 (wild type (WT), and S11A and K168A variants) and Sso2081 (wild type, and S11A and R105A, K106A variants). Catalytic rate constants under single-turnover conditions are plotted. The data points are derived from exponential fits to the triplicate rate measurements presented in Extended Data Fig. 8, with the standard errors derived from curve fitting shown. **c**, Cartoon showing the reaction scheme for conversion of cyclic to linear A_4>P and A_2>P. S11 may participate in the correct positioning of the 2′-OH group of the ribose to facilitate nucleophilic attack, and the basic residue K168 (and R105 and K106 in Sso2081) may stabilize the pentacovalent phosphorus formed in the transition state.

protein family[17]. To confirm that Sso2081 was the nuclease responsible for degradation of cA_4, we expressed and purified the recombinant protein in *Escherichia coli*. The enzyme converted cA_4 to a slower-migrating species on gel electrophoresis, as seen for the activity from *S. solfataricus*, in a reaction independent of divalent metal ions (Fig. 1e). We also noted that the Csx1 nuclease (Sso1389; RCSB Protein Data Bank (PDB) code: 2I71)[14] is found close to an uncharacterized CARF-domain protein of known structure (Sso1393) that is homologous to Sso2081 (Extended Data Figs. 1, 2). Pure recombinant Sso1393 also exhibits cA_4-degradation activity, whereas Csx1 does not (Fig. 1e). By contrast, of the three enzymes only Csx1 displays cA_4-stimulated RNase activity on a linear RNA substrate (Fig. 1f). Sso2081 and Sso1393 degrade cA_4 with single-turnover rate constants of 0.23 ± 0.01 and 0.024 ± 0.0004 min^{-1}, respectively (Fig. 2a and Extended Data Fig. 3). The tenfold-higher specific activity of Sso2081, combined with the higher expression levels of this protein in *S. solfataricus*[18], suggest that it is the major cA_4-degrading enzyme in this organism.

Metal-independent ribonucleases and ribozymes—including the Cas6 nuclease family[19]—share a common mechanism: activation of the 2′-hydroxyl of the ribose sugar as the nucleophile that attacks the phosphodiester bond targeted for cleavage, which leads to products with a 2′,3′- cyclic phosphate and a 5′-hydroxyl moiety[20]. To determine the degradation products of cA_4 treated with Sso2081 and Sso1393, we analysed radioactively labelled species by thin-layer chromatography (Fig. 2b and Extended Data Fig. 4). Sso2081 converted cA_4 into a distinct species (product X) that could be phosphorylated by treatment with polynucleotide kinase, which shows that product X is a linear product with a 5′-OH group (Fig. 2b, lanes 2, 3). For Sso1393, a linear intermediate product (product Y) with a 5′-OH group that converted over time into the final product (product X) was also observed (Fig. 2b, lanes 4–7). Lanes 8 and 9 in Fig. 2b show standards that were generated by cleavage of RNA oligonucleotides by the MazF RNase[14], to generate the species 5′-phospho-diadenylate with a cyclic 2′,3′ phosphate (P-A_2>P, in which '>P' denotes the cyclic phosphate) and 5′-phospho-tetraadenylate (P-A_4>P), respectively. Once the final products of Sso2081 and 1393 are phosphorylated by polynucleotide kinase, they run at the same position as the P-A_2>P standard, whereas the intermediate (product Y) that was observed for Sso1393 runs at the same position as the P-A_4>P standard after a similar treatment with polynucleotide kinase.

To identify the products X and Y, we incubated Sso2081 and 1393 individually with cA_4 and analysed the reaction products by liquid chromatography coupled with high-resolution mass spectrometry (LC–HRMS, Fig. 2c, d). After a 60-min incubation with recombinant Sso2081, the peak corresponding to cA_4 had almost completely disappeared in favour of the main product X with a retention time of 4.4 min, alongside a trace of product Y that eluted at 11.9 min (Fig. 2c and Extended Data Fig. 5). Linear A_2 and A_4 species with a 2′,3′-cyclic phosphate—generated using the MazF toxin[14]—eluted with retention times that were comparable to those of products X and Y, respectively. Mass spectrometry revealed product X to have a neutral mass of 658.104 atomic mass units (AMU), consistent with linear A_2>P. The mass of product Y (1,316.208 AMU), on the other hand, was consistent with A_4>P. The same products were observed for Sso1393 (Fig. 2d and Extended Data Fig. 6). Again, A_2>P (product X) was the main product and A_4>P (product Y) was more visible at early time points, consistent with the slower catalytic rate of Sso1393. Thus, these enzymes break the cA_4 ring using a metal-independent mechanism, which generates a linear A_4>P intermediate and A_2>P products with 5′-OH and

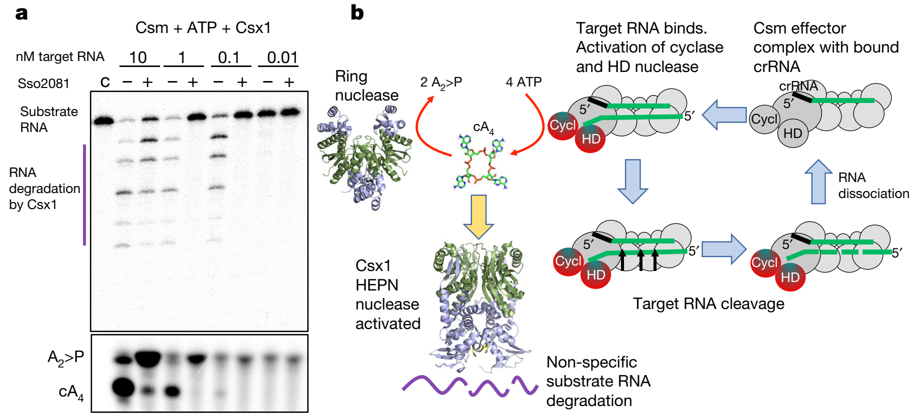


Fig. 4 | Reconstitution of the cA_4 signalling pathway. a, The Csm effector generates cA_4 in proportion to the amount of viral target RNA present[14], activating the HEPN nuclease Csx1. The presence of Sso2081 (2.5 µM) for 1 h before the addition of Csx1 (0.5 µM) for 20 min at 70 °C reversed Csx1 activation partially when 10 nM target RNA was present, and fully when lower amounts of RNA were used. Control (labelled c) shows the experiment in the presence of Sso2081 and absence of target RNA. Levels of cA_4 and A_2>P were monitored by thin-layer chromatography. The data are representative of three separate experiments. **b**, Schematic of the cA_4 signalling system. Type III effectors loaded with CRISPR RNA (crRNA) bind to viral target RNA, activating the HD nuclease—which targets viral DNA—and the cyclase domain. cA_4 is synthesized, which activates the Csx1 RNase. Target RNA cleavage and subsequent dissociation deactivates the cyclase domain, and thus stops synthesis of cA_4. The ring nucleases complete the deactivation of the antiviral state by degrading extant cA_4.

2′,3′-cyclic phosphate termini. We tested the specific activity of Sso2081 against available cyclic nucleotide molecules—cA_6, cA_4, cyclic di-AMP, cyclic di-GMP and cyclic GMP–AMP (Extended Data Fig. 7)—and confirmed that the enzyme is highly specific for the degradation of cA_4, and has a very low level of activity directed against cA_6 and no detectable degradation of the cyclic dinucleotides. We propose that this family of phosphodiesterases be known collectively as ring nucleases.

The structure of Sso1393 (PDB: 3QYF) reveals a canonical CARF domain formed by a homodimeric subunit arrangement, with a C-terminal extension. To elucidate the mechanism of ring nucleases, we docked cA_4 into the CARF domain of Sso1393 (Fig. 3a and Extended Data Fig. 8). This simple model without energy minimization allows hypotheses concerning the enzyme mechanism to be drawn. Using structure-guided multiple sequence alignment of Sso1393, Sso2081 and homologues (Extended Data Fig. 2), we identified conserved residues that are positioned close to the cA_4-binding site that might have a role in catalysis (Fig. 3). We noted the conservation of a lysine (K168 in Sso1393) that is predicted to lie in the centre of the cA_4 binding site, which could have a role in catalysis. Sso2081 has two basic residues (R105 and K106) in this position. Sso1393(K168A) was catalytically inactive, as was the Sso2081(R105A, K106A) variant, which confirms that these residues have an important role in catalysis—possibly by stabilizing the pentacovalent phosphorous generated in the transition state[20] (Fig. 3b and Extended Data Fig. 8). The conserved residue S11 is also implicated in catalysis: the catalytic rate constant was reduced by 3.5- and 32-fold for Sso2081(S11A) and Sso1393(S11A), respectively (Fig. 3b and Extended Data Fig. 8). Consistent with the absence of ring nuclease activity, Csx1 lacks these key active-site residues.

Based on the evidence available, we propose that cleavage of cA_4 is catalysed by binding the molecule in an orientation that enables the 2′-hydroxyl of the ribose to attack the bridging phosphorus, coupled with stabilization of the developing transition state by the basic residue(s) at the base of the binding pocket (Fig. 3c). The single-turnover rate constant for Sso2081 is below 1 min^{-1}, which is comparable to the Cas6 nuclease family (which uses a similar mechanism[19,21]). Given the dimeric organization of the CARF domain, there is the possibility for two active sites to act on opposite sides of the cA_4 ring, consistent with the observation of an A_2>P product. For the slower Sso1393 enzyme, appreciable levels of A_4>P intermediate were observed, which suggests that the two active sites need not function in a concerted manner.

We next wished to reconstitute the cA_4 signalling system in vitro to confirm the function of the ring nucleases. We first tested the ability of ring nucleases to deactivate Csx1, observing that as cA_4 is converted to A_2>P the activation of the Csx1 HEPN RNase is abrogated (Extended Data Fig. 9). We then set up a reaction containing the Csm effector complex, 0.5 mM ATP and a variable concentration of target RNA (from 10 to 0.01 nM), to activate cA_4 synthesis. After incubation for 1 h at 70 °C in the presence or absence of 2.5 µM Sso2081, 0.5 µM Csx1 and radioactively labelled substrate RNA were added. Transcript levels of the *sso2081* gene are about twofold higher than those of the *csx1* gene in uninfected *S. solfataricus* cells[18], and large changes in expression of these genes were not observed on infection with the STIV virus[22]. Target RNA binding by Csm activates the cyclase domain, which switches on cA_4 production. The cyclase domain is deactivated when target RNA is cleaved and dissociates from Csm, leaving the complex free to bind to further targets[14]. Higher concentrations of target RNA thus result in higher levels of cA_4 and stronger activation of Csx1, which degrades a labelled substrate RNA (Fig. 4a). The presence of Sso2081 abrogated the activation of Csx1 in a manner that was dependent on the concentration of target RNA. This recapitulates the situation that is likely to prevail in cells infected with a virus, in which viral RNA load determines the extent of cOA production and speed of subsequent clearance by the ring nuclease (Fig. 4b).

Thus, we have identified a family of ring nucleases containing CARF domains as the enzymes responsible for the degradation of cA_4, which is synthesized by type III CRISPR systems in response to the detection of viral RNA. The ring nucleases act as the off-switch for the system, limiting the damage that is caused by the HEPN ribonucleases—activated by cA_4—once invading RNA has been cleared from the cell. The relatively slow kinetics of cA_4 degradation by ring nucleases are probably consistent with the requirement to maintain the antiviral response for an appropriate period, but we cannot rule out the possibility that other factors increase the cA_4 clearance rate under certain circumstances. Clearly, it will be important to follow up these biochemical studies with genetic analyses to elucidate the function of ring nucleases in the antiviral cOA-signalling pathway in vivo.

Many type III CRISPR systems have multiple associated proteins containing a CARF-domain; some of these proteins may be specialized for the degradation of cOA. Ring nucleases are hard to identify using bioinformatics analysis, as they involve the 'grafting' of a minimal RNase catalytic site onto an extensive cOA binding site in the CARF domain. In a minimal system, it is possible that a single enzyme has a C-terminal HEPN-family RNase coupled to an N-terminal CARF-family ring nuclease. This would allow cOA binding to rapidly switch

on RNase activity and then slowly auto-deactivate by cOA cleavage. However, specialization of CARF proteins as ring nucleases would yield the advantage of enabling levels of cOA-activated RNase and cOA-degrading ring nuclease activity to be controlled independently. It is also possible that cA_6 rings—such as those generated in *Streptococcus thermophilus*—will be processed differently, as they are likely to have more conformational flexibility than cA_4. The recent identification of a wide range of as-yet-uncharacterized CARF-domain proteins across the prokaryotes[8] suggests that we have only scratched the surface of this system. These are fruitful areas for future studies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0557-5.

Received: 1 June 2018; Accepted: 15 August 2018;
Published online 19 September 2018.

1. Mojica, F. J. & Rodriguez-Valera, F. The discovery of CRISPR in archaea and bacteria. *FEBS J.* **283**, 3162–3169 (2016).
2. Makarova, K. S. et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).
3. Jiang, F. & Doudna, J. A. The structural biology of CRISPR–Cas systems. *Curr. Opin. Struct. Biol*. **30**, 100–111 (2015).
4. Niewoehner, O. et al. Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543–548 (2017).
5. Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR–Cas systems. *Science* **357**, 605–609 (2017).
6. Makarova, K. S., Anantharaman, V., Grishin, N. V., Koonin, E. V. & Aravind, L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front. Genet.* **5**, 102 (2014).
7. Burroughs, A. M., Zhang, D., Schäffer, D. E., Iyer, L. M. & Aravind, L. Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. *Nucleic Acids Res.* **43**, 10633–10654 (2015).
8. Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Severinov, K. V. & Koonin, E. V. Systematic prediction of genes functionally linked to CRISPR–Cas systems by gene neighborhood analysis. *Proc. Natl Acad. Sci. USA* **115**, E5307–E5316 (2018).
9. Koonin, E. V. & Makarova, K. S. Discovery of oligonucleotide signaling mediated by CRISPR-associated polymerases solves two puzzles but leaves an enigma. *ACS Chem. Biol.* **13**, 309–312 (2018).
10. Cai, X., Chiu, Y. H. & Chen, Z. J. The cGAS–cGAMP–STING pathway of cytosolic DNA sensing and signaling. *Mol. Cell* **54**, 289–296 (2014).
11. Deng, L., Garrett, R. A., Shah, S. A., Peng, X. & She, Q. A novel interference mechanism by a type IIIB CRISPR–Cmr module in *Sulfolobus*. *Mol. Microbiol.* **87**, 1088–1099 (2013).
12. Hatoum-Aslan, A., Maniv, I., Samai, P. & Marraffini, L. A. Genetic characterization of antiplasmid immunity through a type III-A CRISPR–Cas system. *J. Bacteriol.* **196**, 310–317 (2014).
13. Jiang, W., Samai, P. & Marraffini, L. A. Degradation of phage transcripts by CRISPR-associated RNases enables type III CRISPR–Cas immunity. *Cell* **164**, 710–721 (2016).
14. Rouillon, C., Athukoralage, J. S., Graham, S., Grüschow, S. & White, M. F. Control of cyclic oligoadenylate synthesis in a type III CRISPR system. *eLife* **7**, e36734 (2018).
15. Huynh, T. N. & Woodward, J. J. Too much of a good thing: regulated depletion of c-di-AMP in the bacterial cytoplasm. *Curr. Opin. Microbiol.* **30**, 22–29 (2016).
16. Dey, R. J. et al. Inhibition of innate immune cytosolic surveillance by an *M. tuberculosis* phosphodiesterase. *Nat. Chem. Biol*. **13**, 210–217 (2017).
17. Makarova, K. S. et al. Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).
18. Wurtzel, O. et al. A single-base resolution map of an archaeal transcriptome. *Genome Res.* **20**, 133–141 (2010).
19. Haurwitz, R. E., Sternberg, S. H. & Doudna, J. A. Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *EMBO J.* **31**, 2824–2832 (2012).
20. Yang, W. Nucleases: diversity of structure, function and mechanism. *Q. Rev. Biophys.* **44**, 1–93 (2011).
21. Sokolowski, R. D., Graham, S. & White, M. F. Cas6 specificity and CRISPR RNA loading in a complex CRISPR–Cas system. *Nucleic Acids Res.* **42**, 6532–6541 (2014).
22. Ortmann, A. C. et al. Transcriptome analysis of infection of the archaeon *Sulfolobus solfataricus* with *Sulfolobus* turreted icosahedral virus. *J. Virol.* **82**, 4874–4883 (2008).

Acknowledgements This work was funded by grants from the Biotechnology and Biological Sciences Research Council (REF BB/M000400/1 and BB/M021017/1). M.F.W. is a Wolfson Research Merit Award holder. We acknowledge the contribution of the Mass Spectrometry Unit of the University of St Andrews to this work.

Reviewer information *Nature* thanks E. Koonin, R. Staals and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.S.A. carried out chromatography to identify the major cA_4-degrading enzyme, carried out enzyme assays and analysis, and developed HEPN nuclease deactivation assays; C.R. carried out thin-layer chromatography to identify cA_4 degradation products and reconstituted the cOA signalling pathway; S.Grü. carried out and analysed the mass spectrometry; S.Gra. generated expression plasmids and purified proteins; and M.F.W. oversaw the work, analysed the data and wrote the manuscript. All authors contributed to data analysis and writing.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0557-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0557-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.F.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Purification of cA_4-degrading enzyme from *S. solfataricus* cellular extract. *S. solfataricus* (Sso) P2 was grown as previously described[23] and cells pelleted by centrifugation at 4,000 r.p.m. (Beckman Coulter Avanti JXN-26; JLA8.1 rotor) for 15 min at 4 °C. Cells were suspended in buffer A containing 100 mM sodium phosphate pH 7.0 and 1.5 M ammonium sulfate with one EDTA-free protease inhibitor tablet (Roche). Cells were lysed by sonicating six times for 1 min with 1 min rest intervals on ice, and the lysate was ultracentrifuged at 40,000 r.p.m. (Beckman Coulter Optima L-90K; 70 Ti rotor) and 4 °C for 45 min before filtering and loading onto a phenyl-sepharose column (GE Healthcare) pre-equilibrated with buffer A. Protein was eluted with a linear gradient of buffer B containing 100 mM sodium phosphate pH 7.0 across 10 column volumes (CV). Each fraction was assayed for cA_4-degradation activity and fractions displaying cA_4 degradation were pooled and concentrated using a 10-kDa molecular weight cut-off ultracentrifugal concentrator (Amicon Millipore). Concentrated protein was then further separated by size-exclusion chromatography (S200; 26/60 GE Healthcare) in buffer containing 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) (HEPES) pH 7.5 and 150 mM KCl. Fractions were assayed for cA_4 degradation, and active fractions pooled and concentrated as previously, exchanging into 10 mM 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 6.0 buffer during concentration. Concentrated protein was then loaded on to a 5-ml HiTrap heparin column (GE Healthcare) equilibrated with 10 mM MES pH 6.0 buffer, washing unbound protein with 8% buffer C containing 10 mM MES pH 6.0 and 1.0 M NaCl for 2 CV. Protein was eluted by a linear gradient to 30% buffer C across 5 CV followed by a gradient to 100% buffer C across 4 CV. After assaying for cA_4 degradation, fractions of interest were visualized by SDS–PAGE. A protein band of ~18 kDa—the presence and abundance of which corresponded to a peak in cA_4 degradation—was excised, trypsin-digested and identified by mass spectrometry.

Cloning, expression and purification of Sso2081, Sso1393 and variants. The cloning, expression and purification of Sso1389 (Csx1) has previously been described[14]. A synthetic gene encoding Sso2081 was purchased from Integrated DNA Technologies (IDT), while *sso1393* was PCR amplified from *S. solfataricus* P2 genomic DNA. Genes encoding Sso2081 and Sso1393 were cloned into the pEHisTEV vector[24], and transformed into *E. coli* DH5α competent cells. Sso2081(S11A), Sso2081(R105A, K106A), Sso1393(S11A) and Sso1393(K168A) variants were generated by site-directed mutagenesis using the QuickChange Site-Directed Mutagenesis protocol (Agilent Technologies), with DNA primers purchased from IDT. Sequence-verified constructs were transformed into C43 (DE3) *E. coli* cells for protein expression.

Expression of recombinant Sso2081 and variants was induced with 0.4 mM isopropyl β-D-1-thiogalactopyranoside at an optical density at 600 nm (OD_{600}) of ~0.8, and cells incubated at 16 °C overnight with shaking at 180 r.p.m., before collecting by centrifugation at 4,000 r.p.m. (JLA8.1 rotor) at 4 °C for 15 min. Cells were suspended in lysis buffer (50 mM Tris-HCl pH 8.0, 0.5 M NaCl, 10 mM imidazole and 10% glycerol) with 1 mg ml^{-1} chicken egg lysozyme (Sigma-Aldrich) and one EDTA-free protease inhibitor tablet. Cells were sonicated six times for 1 min with 1 min rest intervals on ice at 4 °C. Cell lysate was then ultracentrifuged at 40,000 r.p.m. (70 Ti rotor) for 45 min at 4 °C, filtered and loaded onto a 5-ml HisTrap FF column (GE Healthcare) equilibrated with buffer D containing 50 mM Tris-HCl pH 8.0, 0.5 M NaCl, 30 mM imidazole and 10% glycerol. After washing unbound protein with 20 CV buffer D, recombinant Sso2081 and mutants were eluted with a linear gradient (0–20%) of buffer D supplemented with 0.5 M imidazole across 15 CV, then holding at 20% for 4 CV. Pooled fractions were concentrated as previously described, and the hexa-histidine affinity tag was removed by incubating protein with tobacco etch virus (TEV) protease (10:1) for 4 h at 37 °C. His-tag-cleaved Sso2081 was isolated from TEV by affinity chromatography as detailed above, eluting with buffer D before further purification by size-exclusion chromatography (S200 26/60; GE Healthcare) in buffer containing 20 mM Tris-HCl pH 8.0, 0.5 M NaCl and 1 mM dithiothreitol (DTT).

Expression of recombinant Sso1393 (wild type and variants) was induced as above, except cells were grown at 16 °C overnight before collection. Cells were lysed and proteins were purified as for Sso2081. All proteins were aliquoted, flash-frozen with liquid nitrogen, and stored at −80 °C.

cA_4 nuclease assays and kinetic analysis. ^{32}P-α-ATP incorporated cA_4 was generated using the Csm complex as previously described[14]. In brief, the Csm complex was incubated for 2 h at 70 °C in 100 μl final reaction volume in a pH 5.5 buffer containing 20 mM MES, 100 mM potassium glutamate and 1 mM DTT supplemented with 2 mM $MgCl_2$, 1 mM ATP, 3 nM ^{32}P-α-ATP and 100 nM *A26* target RNA: (5′-AGGGUCGUUGUUAAGAACGACGUUGUUAGAA GUUGGGUAUGGUGGAGA). The reaction was stopped by phenol–chloroform extraction followed by chloroform extraction. Sso1393, Sso2081 and their mutants were assayed for labelled cA_4 degradation by incubating 2 μM protein dimer with 1/300-diluted, Csm-generated, ^{32}P-labelled cA_4 (0.33 μl per 100 μl reaction) in buffer E (20 mM 2-amino-2-(hydroxymethyl)-1,3-propanediol pH 8.0, 100 mM NaCl, 1 mM EDTA and 1 mM DTT) at 70 °C. At desired time points, a 10-μl aliquot was removed, and the reaction quenched by adding to chilled phenol–chloroform (Ambion). Subsequently, 5 μl of deproteinized reaction product was extracted into 5 μl 100% formamide for denaturing PAGE. Control reactions include cA_4 incubated in buffer E without protein at 4 °C and at 60 °C up to the end point of each experiment (typically 20, 120 or 180 min), phenol–chloroform extracted as above. All experiments were carried out in triplicate. cA_4 degradation was visualized by phosphorimaging following denaturing PAGE (7 M Urea, 20% acrylamide, 1× TBE). For kinetic analysis, cA_4 cleavage was quantified using the Bio-Formats plugin[25] of ImageJ[26] as distributed in the Fiji package[27] and fitted to a single exponential curve using Kaleidagraph (Synergy Software), as previously described[28].

HEPN nuclease deactivation assays. For degradation of cA_4, Csm-generated cA_4 was incubated with 2 μM Sso2081 dimer or 4 μM Sso1393 dimer in buffer E for 60 min and 120 min, respectively. Subsequently, the reaction was deproteinized by phenol–chloroform extraction and diluted twofold in RNase free water. As a control reaction, cA_4 was mock-treated in buffer with water in place of protein and deproteinized as before. Csx1 dimer (500 nM) was incubated with 50 nM *A1* RNA and either no cA_4 activator, 1/100-diluted untreated Csm cA_4, 1/100 mock-treated cA_4 or 1/100 ring-nuclease-treated cA_4 in buffer containing 20 mM MES pH 5.5, 100 mM K-glutamate and 1 mM DTT for 60 min at 70 °C. Reactions were quenched by the addition of a reaction volume equivalent of 100% formamide, and RNA cleavage was assessed by phosphorimaging following denaturing PAGE.

Reconstitution of the cA_4 signalling pathway. The Csm complex (70 nM carrying the CRISPR RNA targeting *A26*) was incubated for 1 h at 70 °C in presence of 2.5 μM of Sso2081 dimer and various concentrations of target *A26* single-strand RNA (ssRNA) in a reaction containing 20 mM MES pH 6.0, 100 mM NaCl, 1 mg ml^{-1} BSA, 2 mM $MgCl_2$, 0.5 mM ATP and a 5′-labelled *A1* ssRNA that is not recognized by the Csm complex (5′-AGGGUAUUAUUUGUUUGUUUCUUCUAAACUA UAAGCUAGUUCUGGAGA). After 1 h incubation, 500 nM of Csx1 dimer was added to the reaction for a further 20 min. Reactions were quenched by deproteination with phenol–chloroform extraction and run on a 20% acrylamide 7 M urea denaturing gel before phosphorimaging to visualize *A1* RNA cleavage. The same experiment was carried out in the presence of 3 pM ^{32}P-α-ATP and unlabelled *A1* RNA to visualize the formation and degradation of cA_4, and 1 μl of the reaction was run on thin-layer chromatography (TLC) before phosphorimaging.

Generation of standards using the MazF nuclease. The *E. coli* toxin–antitoxin MazEF was purified as previously described[14]. Active MazF was liberated either by trypsin (Promega) digestion (1,600:1) at 37 °C for 15 min or by incubation with Factor X Activated (0.1 unit per 1 mg of protein; Sigma-Aldrich) in FXa buffer containing 10 mM Tris-HCl pH 8.0 and 1 mM DTT. For generating linear oligoadenylates A2>P (5′ hydroxyl-ApAp with a 2′,3′-cyclic phosphate) and A4>P, 30 μM *A2* (AAACAUCAG) or *A4* (AAAAACAUCAG) RNA was incubated with MazF in FXa buffer for 1 h at 37 °C. RNA was deproteinized by phenol–chloroform extraction followed by chloroform extraction. For use as standards, A_2>P and A_4>P linear oligoadenylates were 5′-end labelled using ^{32}P-γ-ATP and T4 polynucleotide kinase (PNK; Thermo Fisher Scientific) via its forward reaction.

Thin-layer chromatography. Experiments were performed using silica gel on a 20 × 20 cm TLC plate (Supelco Sigma-Aldrich) containing a fluorescent indicator, allowing visualization of non-radioactive products by ultraviolet (UV) shadowing. Before loading, all samples were deproteinized by phenol–chloroform followed by chloroform extraction. Non-radioactive or radioactive samples (from 0.1 μl to 2 μl) were loaded 1 cm above the bottom of the TLC plate. The TLC plate was then placed in a sealed glass chamber pre-warmed at 30 °C and containing 0.5 cm of a running buffer composed of H_2O (30%), ethanol (70%) and ammonium bicarbonate (0.2 M), pH 9.3. The buffer was allowed to rise along the plate through capillary action until the migration front reached 15 cm. The plate was dried and radioactive sample migration visualized by phosphorimaging while non-radioactive sample migration was pictured under UV light (254 nm).

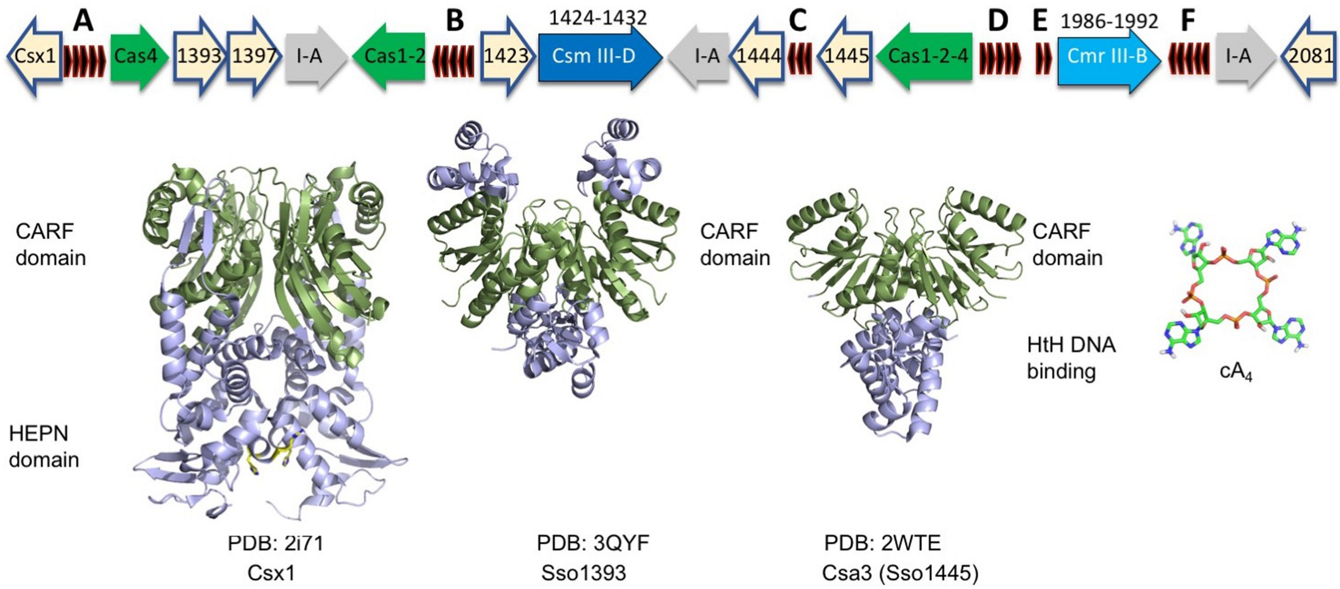
LC–HRMS. LC–HRMS analysis was performed on a Thermo Scientific Velos Pro instrument equipped with HESI source and Dionex UltiMate 3000 chromatography system. Samples were deproteinized as described for TLC. Compounds were separated on a Kinetex 2.6-μm EVO C18 column (2.1 × 100 mm, Phenomenex) using a linear gradient of acetonitrile (B) against 20 mM ammonium bicarbonate (A): 0–5 min 2% B, 5–33 min 2–15% B, 33–35 min 15–98% B, 35–40 min 98% B, 40–41 min 98–2% B, 41–45 min 2% B. The flow rate was 350 μl min^{-1} and the column temperature was 40 °C. UV data were recorded at 254 nm. Mass data were acquired on the FT mass analyser in negative-ion mode with scan range *m*/*z* 150–1,500 at a resolution of 30,000. Source voltage was set to 3.5 kV, capillary temperature was 350 °C, and source heater temperature was 250 °C. Data were analysed using Xcalibur (Thermo Scientific). Extracted ion chromatograms were smoothed using the Boxcar function at default settings.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

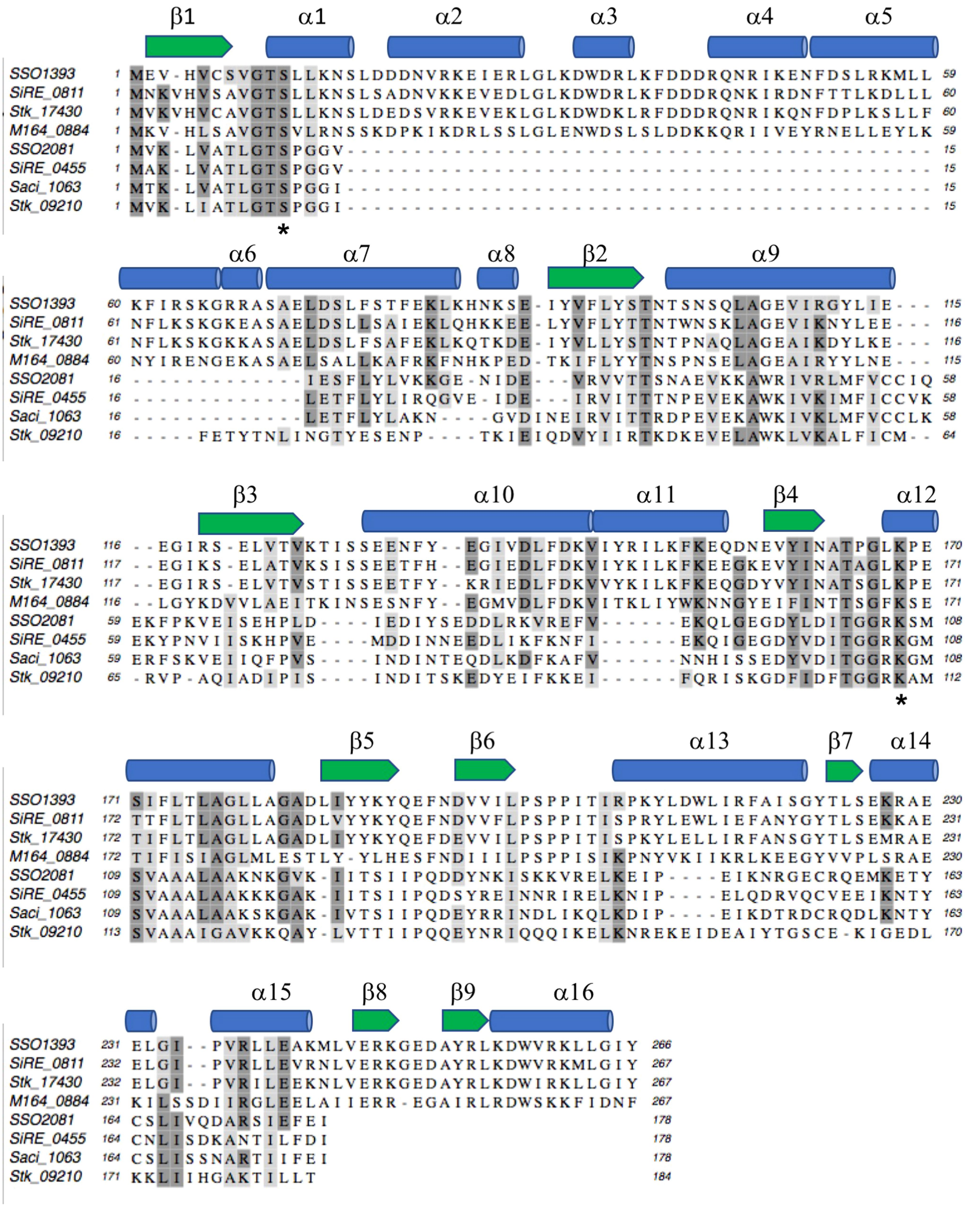
Data availability

All electrophoretic separation and kinetic data are included as Source Data and Supplementary Information. Raw mass spectrometry data are available on request from M.F.W.

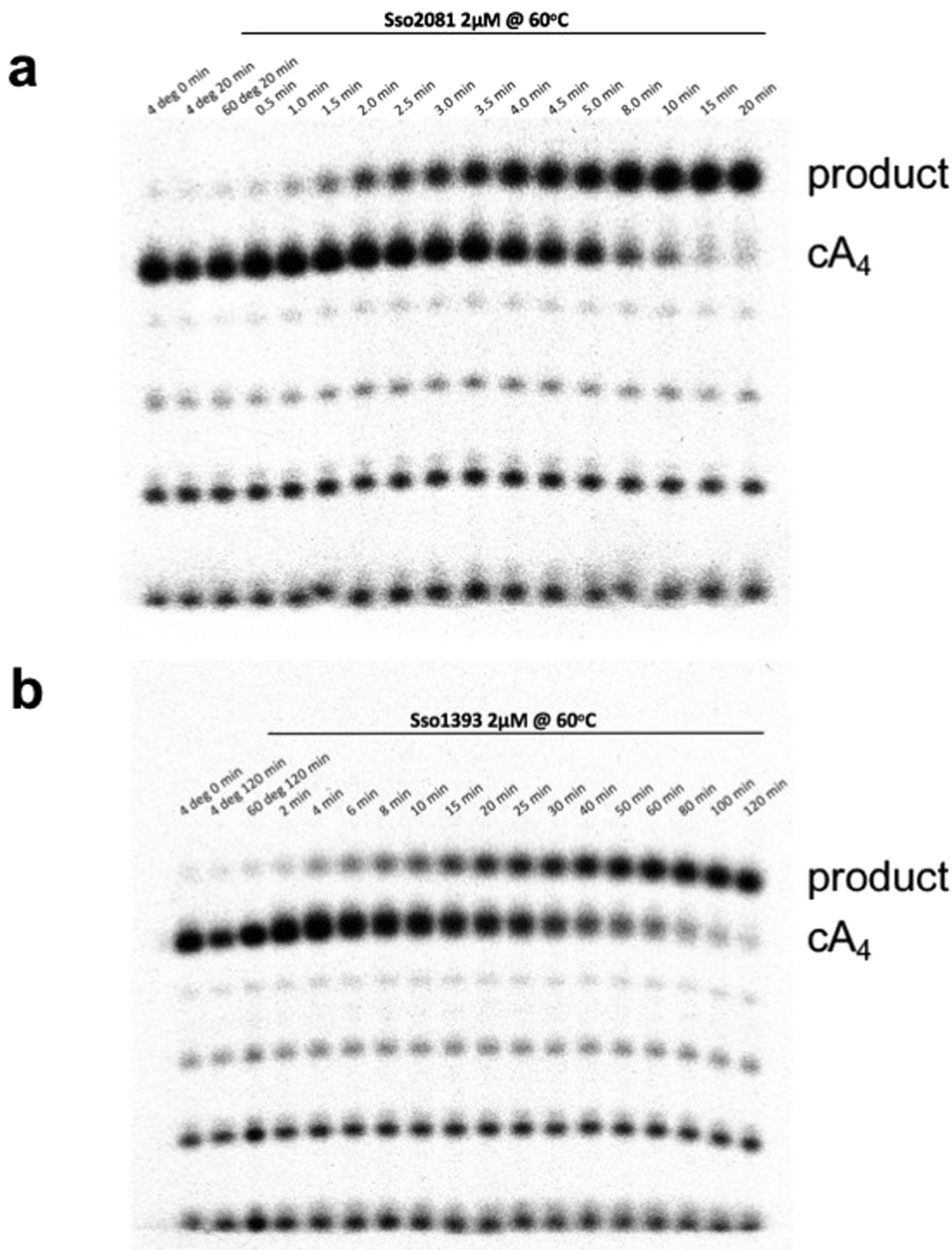
23. Albers, S. V. et al. Production of recombinant and tagged proteins in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Appl. Environ. Microbiol*. **72**, 102–111 (2006).
24. Liu, H. & Naismith, J. H. A simple and efficient expression and purification system using two newly constructed vectors. *Protein Expr. Purif.* **63**, 102–111 (2009).
25. Linkert, M. et al. Metadata matters: access to image data in the real world. *J. Cell Biol*. **189**, 777–782 (2010).
26. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
27. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
28. Sternberg, S. H., Haurwitz, R. E. & Doudna, J. A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* **18**, 661–672 (2012).
29. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc* **4**, 363–371 (2009).



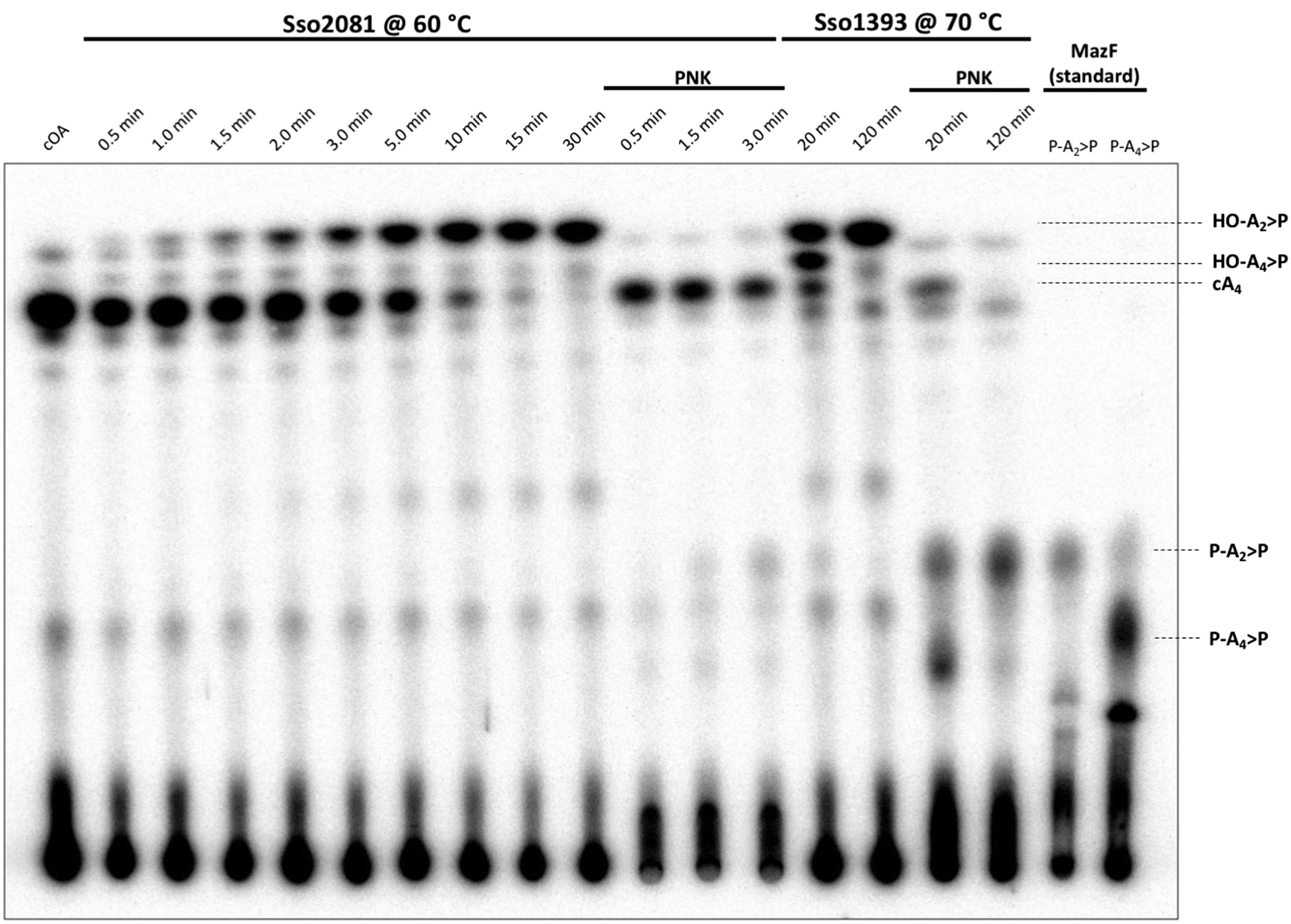
Extended Data Fig. 1 | Genome organization of the CRISPR–Cas locus of *S. solfataricus*. The type I-A, III-B and III-D effector complex operons are depicted, along with genes encoding adaptation proteins and the position of the six CRISPR loci (A–F). Genes outlined in blue encode proteins with CARF domains. The structures of three CARF family proteins are shown with CARF domains coloured green, along with the structure of cA_4.



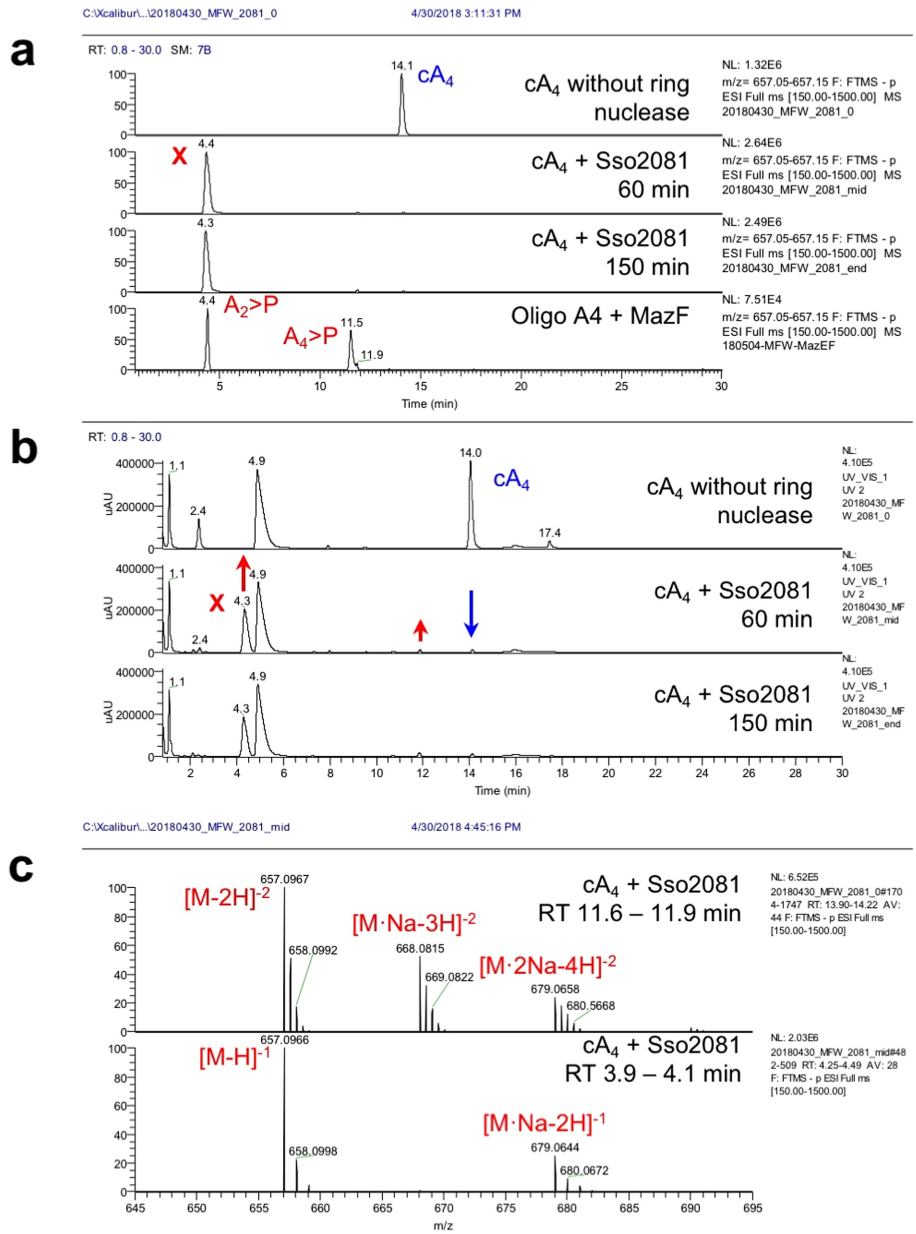
Extended Data Fig. 2 | Structure-guided sequence alignment of Sso1393, Sso2081 and orthologues. Multiple sequence alignment showing Sso1393 together with three homologues, aligned with Sso2081 and three homologues. Secondary structure is shown above the alignment, based on the structure of Sso1393 (PDB: 3QYF). S11 and K168 are indicated by asterisks. Conserved residues are shaded. Sequences aligned are from *S. solfataricus* (Sso1393, Sso2081); *Sulfolobus islandicus* REY15A (SiRE_0811, SiRE_0455); *S. islandicus* M.16.4 (M164_0884); *Sulfolobus acidocaldarius* (Saci_1063) and *Sulfolobus tokodaii* (Stk_17430).



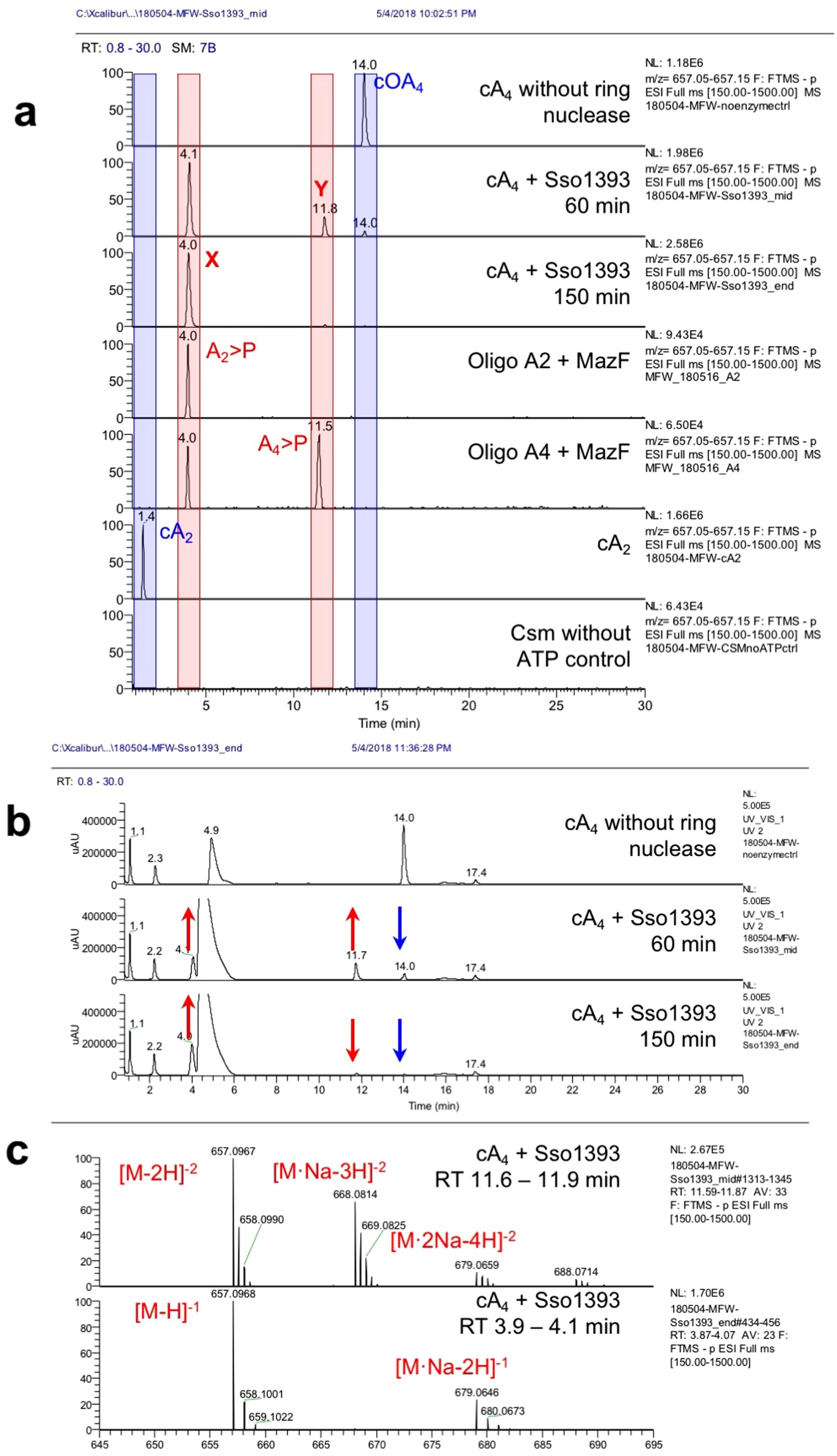
Extended Data Fig. 3 | Single-turnover kinetic analysis of cA_4 degradation by Sso2081 and Sso1393. a, **b**, Representative phosphorimages of PAGE, analysing the reaction of radioactively labelled cA_4 with 2 μM Sso2081 (**a**) or 2 μM Sso1393 (**b**) at 60 °C over time. Experiments were carried out in triplicate, quantified by phosphorimaging and the fraction of cA_4 cleaved is plotted in Fig. 3b.



Extended Data Fig. 4 | Sso2081 and Sso1393 cA_4-degradation mechanism investigated by TLC. Csm-generated cOA (lane 1) was incubated with 2 μM Sso2081 dimer at 60 °C to determine the intermediate (Y) and final (X) reaction product over time (lanes 2–10). Lanes 11–13 show reaction product seen in lanes 2, 4 and 6, 5′-end phosphorylated using T4 PNK for identification of reaction intermediates and products by comparison to the 5′-end phosphorylated HO-A_2>P and HO-A_4>P standards generated with MazF nuclease. Lane 14 and lane 15 show the reaction products of 2 μM Sso1393 dimer incubated with cA_4 at 70 °C for 20 and 120 min, respectively. Reaction product from lanes 14 and 15 are 5′-end phosphorylated by PNK for comparison with P-A_2>P and P-A_4>P standards. Comparison of PNK-treated reaction product to standards showed the presence of a low amount of intermediate (P-Y) during the Sso2081 cA_4-cleavage reaction, which migrated similarly to the P-A_4>P standard and did not change in abundance over time, whereas the abundance of the final product (P-X) increased over time. By contrast, comparison of Sso1393 PNK-treated reaction products at 20 min and 120 min showed a decrease of the intermediate (P-Y) over time and an increase of product (P-X).

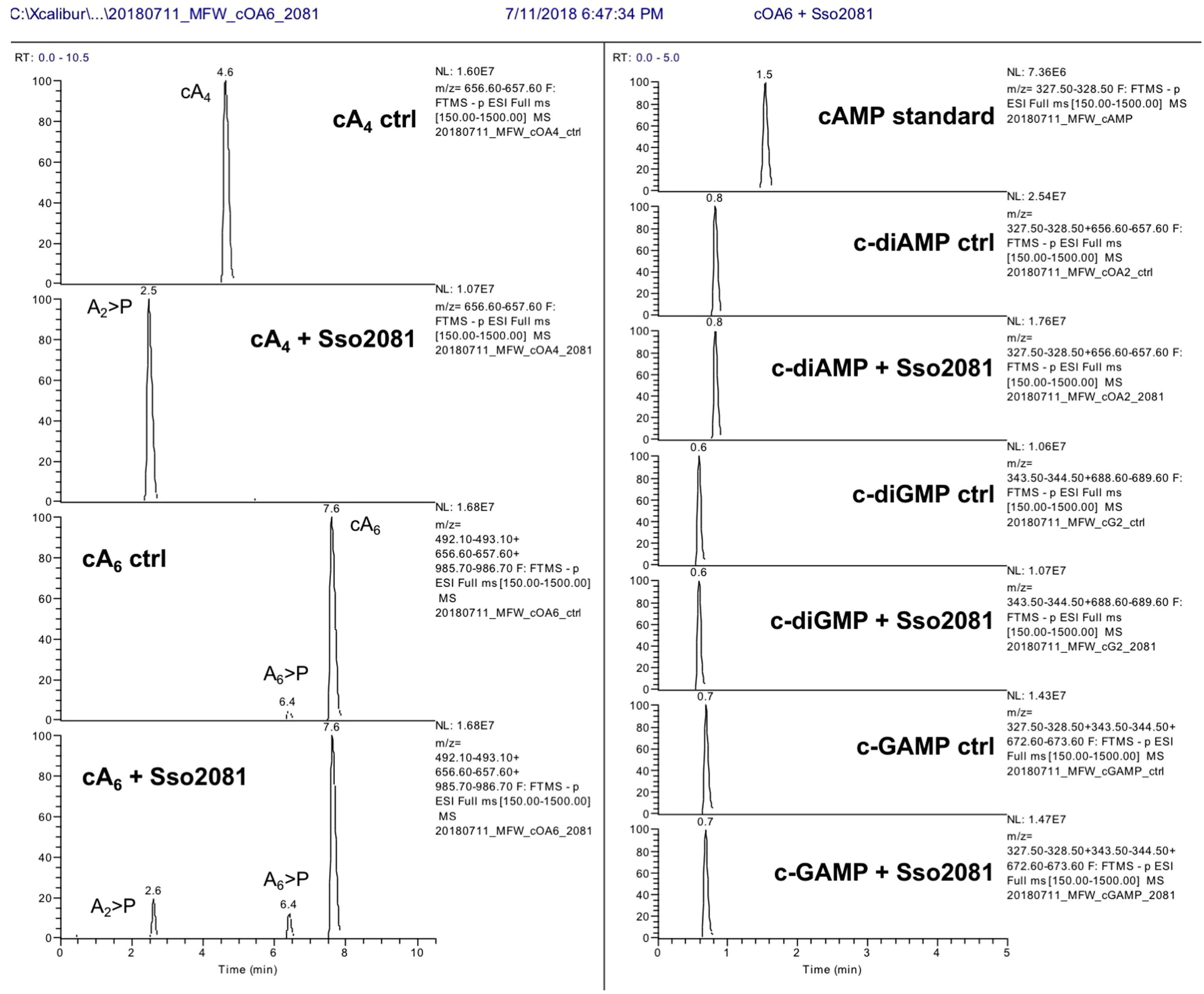


Extended Data Fig. 5 | Liquid chromatography–mass spectrometry analysis of Sso2081 reactions. a, Ion chromatograms extracted for m/z 657.1 (cA_2^{-1}, $A_2{>}P^{-1}$, cA_4^{-2} and $A_4{>}P^{-2}$). Individual lanes are labelled. cOAs (mainly cA$_4$), derived from reaction of Csm with ATP, were incubated with Sso2081 for 60 and 150 min. Linear oligoadenylates with 2′,3′-cyclic phosphate were derived from hydrolysis of *A4* RNA oligonucleotide with MazF. **b**, UV traces at 254 nm. Peaks that change in intensity over the course of the enzymatic reaction are indicated by arrows. The three peaks that decreased or increased over the course of the reaction all match the changes in abundance of the m/z 657.1 species. No changes are observed after 60 min of reaction time. The broad peak at 4.9 min is an unknown contaminant that probably resulted from the phenol–chloroform extraction. Shifts in retention time are possibly due to matrix effects. **c**, Mass spectra of cA$_4$ and product X. Calculated for $C_{20}H_{23}N_{10}O_{12}P_2^{-1}$ (cA$_2$/A$_2$>P) m/z 657.0978, found 657.0966 (δm 1.8 ppm); calculated for $C_{40}H_{46}N_{20}O_{24}P_4^{-2}$ (cA$_4$/A$_4$>P) m/z 657.0978, found 657.0967 (δm 1.6 ppm). The data presented are representative of experiments performed in duplicate.

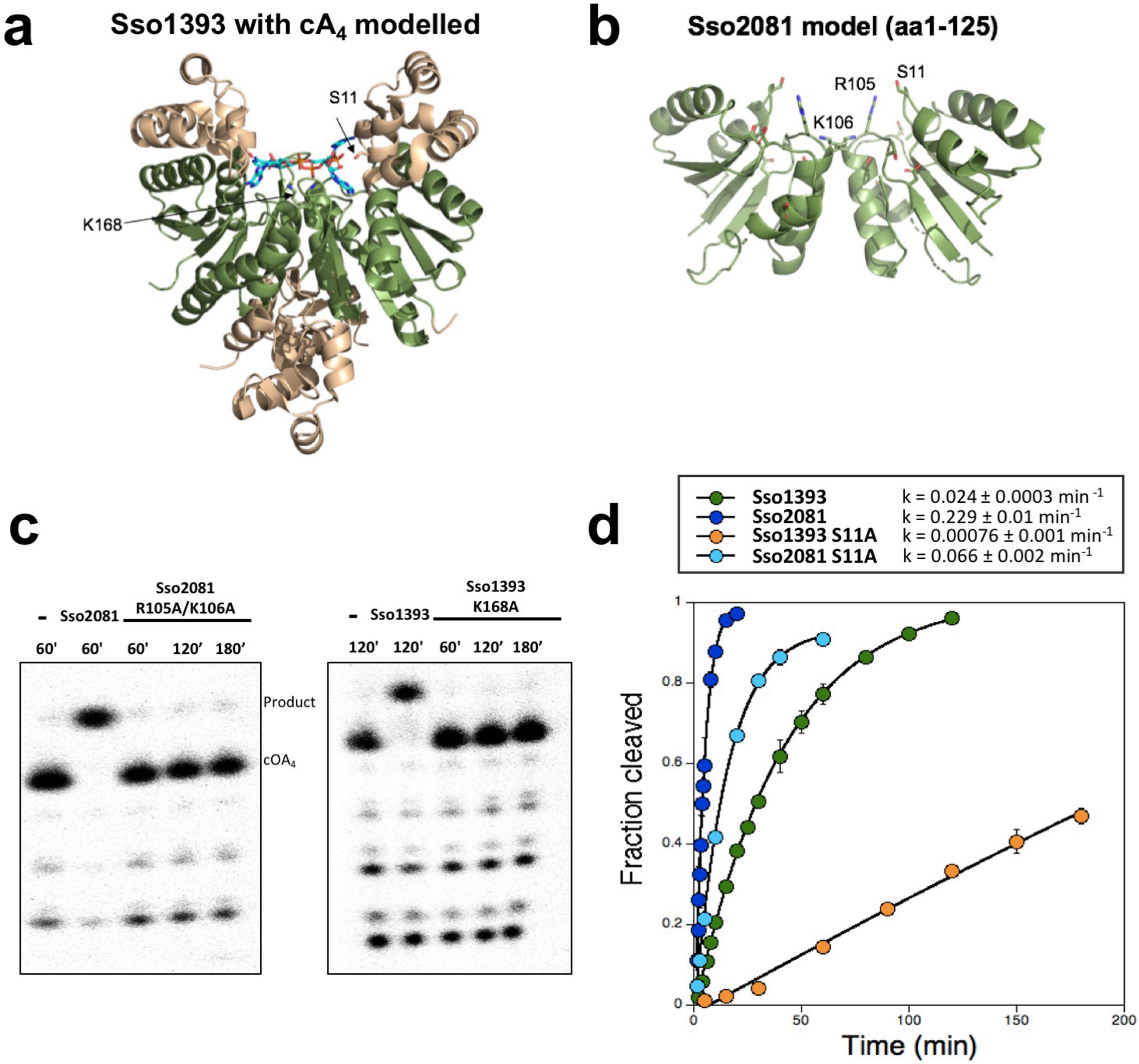


Extended Data Fig. 6 | See next page for caption.

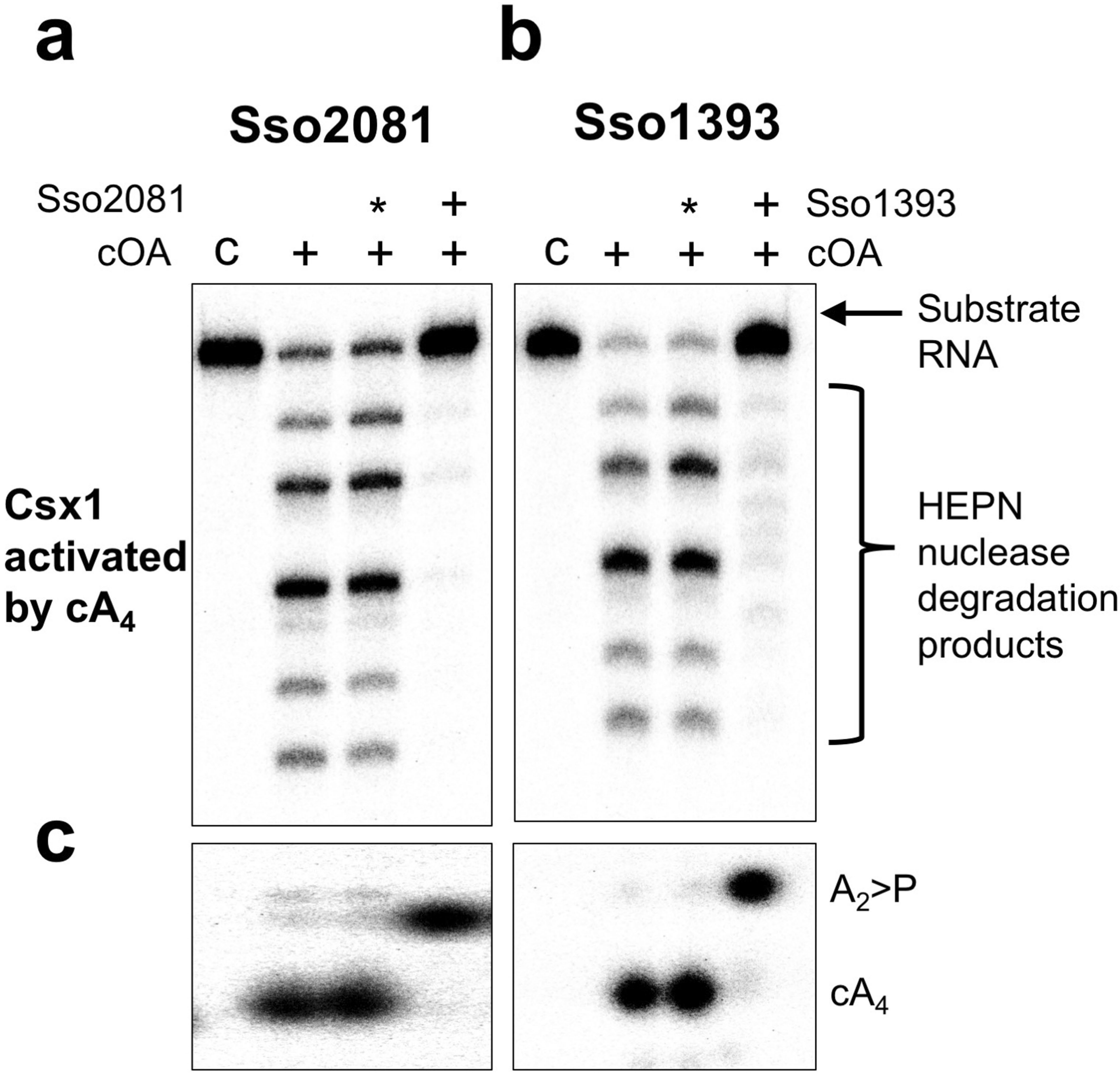
Extended Data Fig. 6 | Liquid chromatography–mass spectrometry analysis of Sso1393 reactions. a, Ion chromatograms extracted for *m*/*z* 657.1 (cA_2^{-1}, $A_2{>}P^{-1}$, cA_4^{-2} and $A_4{>}P^{-2}$). Individual lanes are labelled. cOAs (mainly cA_4), derived from reaction of Csm with ATP, were incubated with Sso1393 for 60 and 150 min, or without ring nuclease for 150 min. A control in which the ATP from the Csm cyclase reaction had been omitted was also analysed as control. Linear oligoadenylates with 2′,3′-cyclic phosphate were derived from hydrolysis of suitable DNA oligonucleotide substrates with the toxin MazF; cA_2 was a commercially available standard. The traces show clearly the difference in retention time between the linear and cyclic isomers. **b**, UV traces at 254 nm. Peaks that change in intensity over the course of the enzymatic reaction are indicated by arrows. The three peaks that decreased or increased over the course of the reaction all match the changes in abundance of the *m*/*z* 657.1 species. The broad peak at 4.1–6 min is an unknown contaminant that probably resulted from the phenol–chloroform extraction. **c**, Mass spectra of species X and Y. Calculated for $C_{20}H_{23}N_{10}O_{12}P_2^{-1}$ (cA_2/$A_2{>}P$) *m*/*z* 657.0978, found 657.0968 (δm 1.4 ppm); calculated for $C_{40}H_{46}N_{20}O_{24}P_4^{-2}$ (cA_4/$A_4{>}P$) *m*/*z* 657.0978, found 657.0967 (δm 1.6 ppm). The data presented are representative of experiments performed in duplicate.



Extended Data Fig. 7 | Specificity of Sso2081 for cyclic nucleotide substrates. Various cyclic di- and oligonucleotides were incubated in the absence (‘ctrl’) or presence of Sso2081 as described in Methods. Dinucleotides and cA_6 were obtained from BIOLOG Life Science Institute (Bremen), and cA_4 was obtained from an enzymatic reaction with *S. solfataricus* Csm. Protein-free extracts were analysed by LC–HRMS essentially as for Fig. 2, but using a shorter column and gradient. Extracted-ion chromatograms for substrate and expected products are shown in each panel. No reaction was observed for any of the cyclic dinucleotides. cA_4 was completely converted to linear A_2>P by Sso2081, whereas only a small percentage of cA_6 was converted. This demonstrates a clear preference of Sso2081 for its physiological substrate, cA_4. The data presented are representative of experiments performed in duplicate.



Extended Data Fig. 8 | Structure and kinetic analysis of Sso1393 and Sso2081. a, Structure of Sso1393 (PDB: 3QYF), with cA_4 docked at the active site. This is an orthogonal view to that shown in Fig. 3. The CARF domain is coloured green. The side chains of S11 and K168 are shown. **b**, Model of the Sso2081 structure. Only the CARF domain (amino acids 1–125) is modelled. Conserved residues S11, R105 and K106 are labelled. Model generated using Phyre[29]. **c**, Representative phosphorimages of denaturing PAGE, assessing the cA_4-degradation activity of Sso2081 compared to its catalytically inactive Sso2081(R105A, K106A) mutant over time (left), and Sso1393 activity compared to Sso1393(K168A) mutant (right). All reactions were carried out at 70 °C with 2 μM protein dimer. The data presented are representative of experiments performed in triplicate. **d**, Single-turnover kinetics of Sso2081 and Sso1393 plotted alongside their active-site mutants (Sso2081(S11A) and Sso1393 (S11A)) and fitted to an exponential equation. Rate constants are displayed with the legend (Sso2081, 0.23 ± 0.01 min^{-1}; Sso2081(S11A), 0.066 ± 0.002 min^{-1}; Sso1393, 0.024 ± 0.0003 min^{-1} and Sso1393(S11A), 0.00076 ± 0.001 min^{-1}). Experiments were carried out in triplicate with means and standard deviation shown. The data presented are technical replicates and are representative of experiments performed in duplicate.



Extended Data Fig. 9 | Ring nucleases abrogate activation of the Csx1 nuclease by converting cA_4 to A_2>P. a, In the absence of cA_4, Csx1 had no RNase activity (lane 'c'), but addition of cA_4 resulted in degradation of the substrate RNA molecule. When cA_4 was pre-incubated for 1 h at 70 °C with Sso2081 in buffer E before addition to the assay, Csx1 RNase activity was abolished. Mock treatment using buffer instead of Sso2081 (denoted '*') had no effect. **b**, As for **a**, but pre-incubated with Sso1393 for 2 h at 70 °C. **c**, Using radioactive cA_4, the deactivation of Csx1 by Sso2081 and Sso1393 was observed to correlate with the conversion of cA_4 into A_2>P. Results shown are representative of experiments performed in at least triplicate.

https://doi.org/10.1038/s41586-018-0567-3

Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing

Binbin Lai[1], Weiwu Gao[1,2], Kairong Cui[1], Wanli Xie[1,3], Qingsong Tang[1], Wenfei Jin[4], Gangqing Hu[1], Bing Ni[2] & Keji Zhao[1]*

Nucleosome positioning is critical to chromatin accessibility and is associated with gene expression programs in cells[1–3]. Previous nucleosome mapping methods assemble profiles from cell populations and reveal a cell-averaged pattern: nucleosomes are positioned and form a phased array that surrounds the transcription start sites of active genes[3–6] and DNase I hypersensitive sites[7]. However, even in a homogenous population of cells, cells exhibit heterogeneity in expression in response to active signalling[8,9] that may be related to heterogeneity in chromatin accessibility[10–12]. Here we report a technique, termed single-cell micrococcal nuclease sequencing (scMNase-seq), that can be used to simultaneously measure genome-wide nucleosome positioning and chromatin accessibility in single cells. Application of scMNase-seq to NIH3T3 cells, mouse primary naive CD4 T cells and mouse embryonic stem cells reveals two principles of nucleosome organization: first, nucleosomes in heterochromatin regions, or that surround the transcription start sites of silent genes, show large variation in positioning across different cells but are highly uniformly spaced along the nucleosome array; and second, nucleosomes that surround the transcription start sites of active genes and DNase I hypersensitive sites show little variation in positioning across different cells but are relatively heterogeneously spaced along the nucleosome array. We found a bimodal distribution of nucleosome spacing at DNase I hypersensitive sites, which corresponds to inaccessible and accessible states and is associated with nucleosome variation and variation in accessibility across cells. Nucleosome variation is smaller within single cells than across cells, and smaller within the same cell type than across cell types. A large fraction of naive CD4 T cells and mouse embryonic stem cells shows depleted nucleosome occupancy at the de novo enhancers detected in their respective differentiated lineages, revealing the existence of cells primed for differentiation to specific lineages in undifferentiated cell populations.

To understand the principles that underline chromatin heterogeneity as related to nucleosome positioning and chromatin accessibility, we developed the scMNase-seq technique to simultaneously measure nucleosome positioning and chromatin accessibility in single cells. We applied scMNase-seq to 48 NIH3T3 single cells, 198 mouse embryonic stem cells (ESCs), and 278 mouse naive CD4 T cells, obtaining on average about 3, 0.9 and 0.7 million unique fragments, respectively, for each cell type (Fig. 1a, Supplementary Table 1). Sequence reads from sorted human or mouse cells from a mixed population mapped exclusively to the respective genome, suggesting that there was no DNA contamination across cells (Extended Data Fig. 1a). Pooled single-cell reads revealed a size distribution that was consistent with that obtained by bulk-cell MNase-seq (Extended Data Fig. 1b). We considered fragments with a length between 140 and 180 bp as canonical nucleosomes, and fragments with a length $\leq$ 80 bp as subnucleosome-sized particles (Extended Data Fig. 1b, c). Compared to CD4 T cells and mouse ESCs (Fig. 1b), NIH3T3 libraries have the largest number of non-redundant reads (Extended Data Fig. 1d) and the highest genomic coverage (5–30%) of nucleosomes—probably owing to the polyploidy of NIH3T3 cells (Extended Data Fig. 1e). Nevertheless, all three cell types have a similar nucleosome density across different genomic regions, which suggests that representation of the genome is relatively even (Extended Data Fig. 1f). The nucleosome positioning and the enrichment of subnucleosome-sized particles surrounding DNase I hypersensitive sites (DHSs), the transcription start sites (TSSs) of active genes and CTCF-binding sites were consistent between pooled scMNase-seq and bulk-cell MNase-seq data (Fig. 1c, Extended Data Fig. 2a–h). The density of subnucleosome-sized particles from pooled single cells is correlated with the DNase I tag density at DHSs and with gene expression at TSSs, suggesting that subnucleosome-sized particles are predictive of chromatin accessibility (Extended Data Fig. 2i, j). Moreover, the percentage of DHSs detected by scMNase-seq was higher than that detected by single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq)[10] with the same sequencing redundancy (owing to the higher complexity and non-redundant read-number of scMNase-seq libraries), although when using scMNase-seq the percentage of recovered DHSs per non-redundant read for subnucleosome-sized particles was relatively lower than scATAC-seq fragments (Extended Data Fig. 2k–n). Nucleosome positions from single cells, aggregated nucleosome density from pooled single cells and tag density from bulk-cell MNase-seq at representative cell-type-specific genes are shown for all three cell types (Fig. 1d). Notably, the similarity of aggregated nucleosome profiles between pooled single cells and bulk cells is correlated with nucleosome positioning stringency and nucleosome coverage, and is higher for active promoters than it is for silent promoters (Fig. 1d, Extended Data Fig. 2o). These results demonstrate that scMNase-seq can simultaneously measure nucleosome positioning and chromatin accessibility in single cells.

Although nucleosome positioning[13] is well-studied[5,14–16] on the basis of large numbers of pooled cells, genome-wide nucleosome spacing patterns are poorly understood because current knowledge about nucleosome spacing is limited to the positioned nucleosomes[7,17,18]. We profiled the distribution of nucleosome-to-nucleosome distance for different single cells and used relative peak height to measure the uniformity of nucleosome spacing for both positioned and non-positioned nucleosomes (Extended Data Fig. 3a–c, Supplementary Methods). This analysis revealed a high degree in spacing uniformity in single cells regardless of positioning stringency; decreased uniformity in spacing was observed as positioning stringency decreased, when using either the pooled single cells or bulk-cell MNase-seq data[19] (Extended Data Fig. 3d). The bulk-cell MNase-seq data failed to reveal the actual spacing pattern owing to the mixture of non-positioned nucleosomes from a population of different cells.

The degree of uniformity in spacing in the promoter regions of silent genes is higher than that of active genes (Fig. 2a, b, Extended Data Fig. 4a, b), and uniformity is higher in non-DHS than in DHS regions (Fig. 2c, d, Extended Data Fig. 4a–c). Notably, the higher uniformity of spacing in non-DHS regions was also observed in single haploid mouse ESCs and haploid chromosome X in single mouse ESCs (Extended Data

[1]Laboratory of Epigenome Biology, Systems Biology Center, National Heart, Lung and Blood Institute, NIH, Bethesda, MD, USA. [2]Department of Pathophysiology and High Altitude Pathology, Third Military Medical University, Chongqing, China. [3]Institute of Anesthesiology and Critical Care Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [4]Department of Biology, South University of Science and Technology of China, Shenzhen, China. *e-mail: zhaok@nhlbi.nih.gov

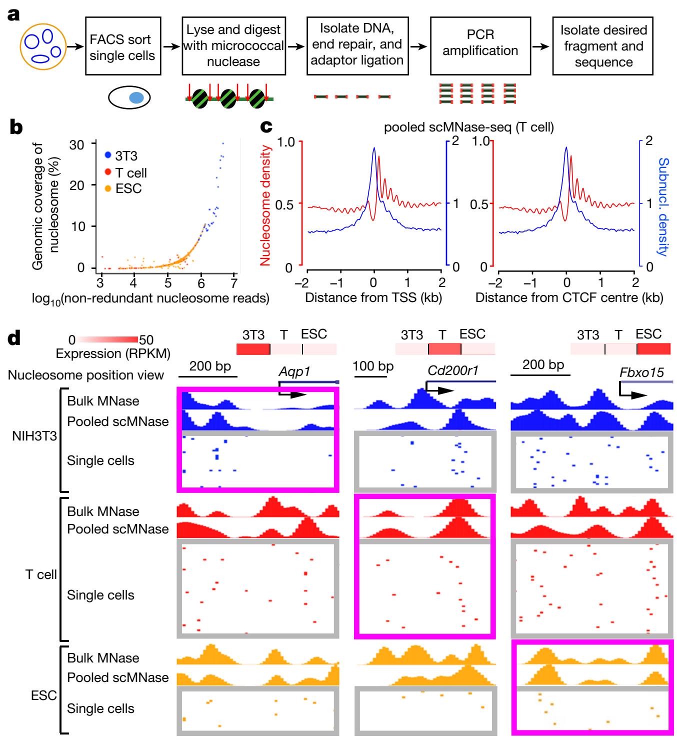


Fig. 1 | scMNase-seq simultaneously measures the positions of nucleosomes and subnucleosome-sized particles in single cells. **a**, Schema of scMNase-seq. **b**, Plot of non-redundant nucleosome read number (*x* axis) and genomic coverage of nucleosomes (*y* axis) for single NIH3T3 cells, CD4 T cells and mouse embryonic stem cells. **c**, Average density profiles of nucleosomes (red) and subnucleosome-sized particles (blue) relative to TSS of active genes (left) and CTCF-binding sites (right) for pooled CD4 T cells scMNase-seq data. Subnucl., subnucleosome-sized particles. **d**, Genome browser view of single-cell nucleosome positions for NIH3T3 cells, CD4 T cells and mouse ESCs at TSSs of three representative cell-type-specific gene loci. Single-cell libraries that have at least one nucleosome within any of three genomic regions are shown. Tracks for tag density of corresponding bulk-cell MNase-seq data (one representative from two repeated experiments is shown) and pooled scMNase-seq data (all single cell libraries with detected nucleosomes in selected genomic regions are included) are also shown. The nucleosome maps at expressed genes for each cell type are highlighted with pink rectangle. The expression levels of genes are shown in the heat map above the tracks. 3T3, NIH3T3 cells; T, CD4 T cells; ESC, mouse ESC; RPKM, reads per kilobase of transcript per million mapped reads.

Fig. 4d–g), and was independent of MNase concentration (Extended Data Fig. 4h–m). Furthermore, nucleosome spacing in active chromatin regions associated with H3K4me1, H3K4me3, H3K27ac, H3K9ac and H2AZ shows a lower degree of uniformity than transcribed regions marked by H3K36me3, heterochromatic regions marked by H3K27me3 or not marked by any of the histone modifications that we studied (Extended Data Fig. 4n, o).

We next measured variation in nucleosome positioning not only across cells but also within single cells (across different alleles) by calculating the mean value of distances between two overlapping nucleosomes within genomic regions related to a particular feature (for example, active promoters) (Extended Data Fig. 5a). As expected, variation in nucleosome positioning around the TSSs of active genes—where nucleosomes are phased relative to TSS—is smaller than that around the TSSs of silent genes (Fig. 2e). In addition, nucleosome positions show smaller variation at the centre of DHSs and the centre of chromatin regions enriched in active histone modifications than they do elsewhere (Extended Data Fig. 5b–g).

The results above reveal that there are different rules of nucleosome organization in different chromatin regions. In silent chromatin states—such as in repressed promoters and heterochromatic regions—nucleosomes are highly uniformly spaced, but are not positioned relative to the underlying genomic DNA across different arrays. By contrast, in active chromatin states—such as transcribed promoters and DHS regions—nucleosomes are positioned but are not as uniformly spaced (Fig. 2f). This model was further supported by the observation that nucleosomes in promoter regions of silent genes, non-DHS regions and heterochromatic regions show higher synchronized shift scores than nucleosomes in promoters of active genes, DHS regions and regions marked by active histone modifications (Extended Data Fig. 6a, b). Furthermore, the synchronized shift score is dependent on nucleosome spacing; the highest scores are in the spacing range of 180–185 bp, which is dominant throughout the genome in all single cells (Extended Data Fig. 6c, d). The nucleosome spacing might indicate a stable structure for packaging nucleosomes[20] in silent chromatin states, which is probably collectively determined by chromatin assembly factors[21], linker histones[22,23] and the environment surrounding chromatin fibres. In active states, ATP-dependent chromatin remodelling activities[15,24] may reposition nucleosomes[1,25] and consequently change the local nucleosome spacing to facilitate chromatin accessibility and gene transcription. Notably, the average nucleosome spacing surrounding the DHSs is shorter than that in non-DHS regions (Extended Data Fig. 6e, f), which may be the result of repositioning of the nucleosomes to allow accessibility of the DHS regions.

Although nucleosomes are positioned surrounding DHSs to ensure chromatin accessibility[7], extensive heterogeneity of chromatin accessibility across different single cells[10,12] implies heterogeneity of nucleosome positioning at the same DHS. Profiling nucleosome-to-nucleosome distances over DHSs reveals two distinct peak patterns: one has a summit at about 190 bp and the other has a summit at about 300 bp, which presumably corresponds to two different chromatin states (closed or open) (Fig. 3a, b). More than 80% of the DHSs have both spacing types at the same DHS in different single cells (Fig. 3c), and the higher DNase I tag density at DHSs measured in bulk cells[26] is associated with more wide-spacing DHSs in single cells (Extended Data Fig. 7a). Furthermore, the DHSs with a higher fraction of wide space—which is not related to MNase digestion—are associated with higher accessibility, when measured by bulk-cell DNase I hypersensitive sites sequencing (DNase-seq) or by scMNase-seq subnucleosome-sized particles (Extended Data Fig. 7b–d), and with lower variation in DHS accessibility and nucleosome positioning across different single cells (Fig. 3d, e). These results indicate that one DHS may have two types of nucleosome organization (wide or narrow spacing) across different single cells; the degree of accessibility of a DHS as well as the variation in DHS accessibility and nucleosome positioning across cells are directly linked to the ratio between the two states of nucleosome organization in different single cells.

Furthermore, variation in nucleosome positioning around DHSs is positively correlated with variation in accessibility across different single cells (Fig. 3f). The fraction of single cells with nucleosomes positioned around DHSs is correlated with the number of cells detected as DHSs (Extended Data Fig. 7e). The variation in nucleosome positioning around TSSs in different single cells is also correlated with variation in gene expression. The TSSs with +1 nucleosomes that show higher variation in nucleosome positioning also show higher variation in expression across different single cells (Fig. 3g). Genes for which expression was detected in a higher fraction of single cells exhibit positioned +1 nucleosomes in a higher fraction of single cells than do the genes with a lower fraction of expression (Extended Data Fig. 7f). The top 1,000 active genes with smallest nucleosome variance around their TSS across cells are enriched in common biological processes such as translation and protein transport (Extended Data Fig. 7g), consistent with the notion that house-keeping genes display less variation in nucleosome positioning. Furthermore, variation within a cell in nucleosome positioning around DHSs, or at the +1 nucleosome of the TSSs of active genes, is smaller than that across different single cells (Extended Data Fig. 7h, i). The variation in nucleosome positioning within the cell type is smaller than that across different cell types

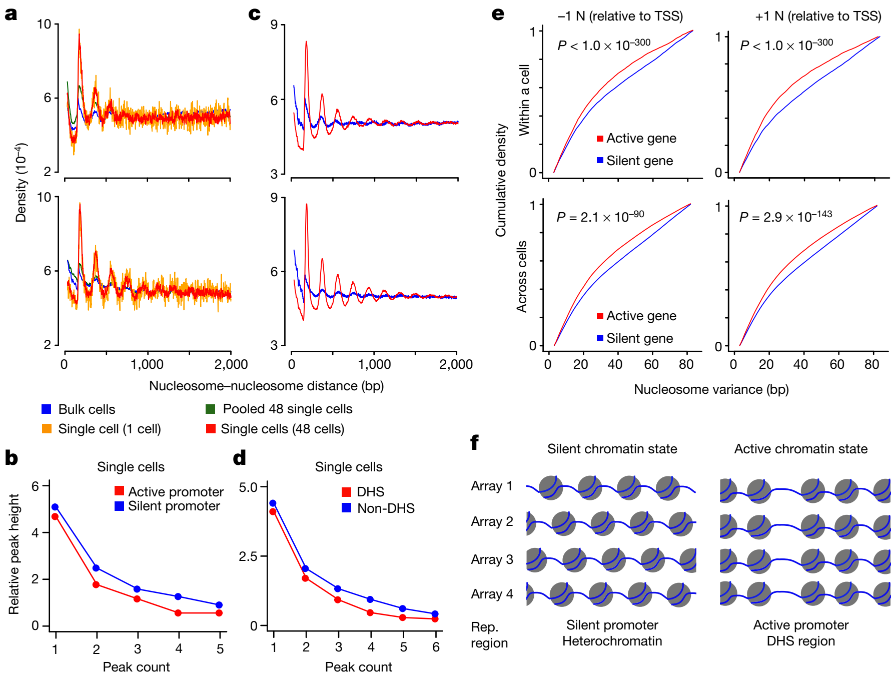


Fig. 2 | Profiling nucleosome positioning and spacing in single cells reveals distinct nucleosome organization principles at active and silent chromatin regions. a, Density plots of nucleosome-to-nucleosome distance within active-gene promoters (top) and silent-gene promoters (bottom) for bulk-cell MNase-seq, pooled 48 NIH3T3 single cells, one representative single cell and 48 single-cell scMNase-seq datasets. **b**, The relative peak heights based on the data from **a** reveal a higher degree of uniformity in spacing within silent-gene promoters than active-gene promoters. **c**, Density plots of nucleosome-to-nucleosome distance within DHS regions (top) and non-DHS regions (bottom) for bulk-cell MNase-seq (blue) and 48 single-cell scMNase-seq (red) datasets. **d**, The relative peak heights based on the data from **c** reveal a higher degree of uniformity in spacing within non-DHS regions than DHS regions. **e**, Cumulative density of variance in nucleosome positioning in active and silent genes within a cell (top) and across single cells (bottom), at −1 (left) and +1 (right) nucleosomes relative to the TSS. Top left, $n = 7{,}574$ and 13,107 nucleosome pairs for active and silent genes, respectively; bottom left, $n = 164{,}512$ and 304,847 nucleosome pairs for active and silent genes, respectively; top right, $n = 11{,}388$ and 17,631 nucleosome pairs for active and silent genes, respectively; bottom right, $n = 237{,}006$ and 416,328 nucleosome pairs for active and silent genes, respectively. *P* values were calculated using one-sided Mann–Whitney *U*-test. **f**, Cartoon illustrating nucleosome organization patterns in silent (left) and active (right) chromatin states. Rep., representative genomic region.

(Extended Data Fig. 7j). Clustering based on similarity in nucleosome positioning at all the DHSs across all the single cells from three cell types separated these single cells into three clusters that correspond to the respective cell types—this clustering is independent of experiment time and fragment-size ratio (Extended Data Fig. 7k).

The DNA sequence has an important role in nucleosome positioning[4,14,16]. Consistent with a previous report[14], we observed high CC, GG and GC frequency in nucleosome-occupied sequences and high AA, TT, AT and TA frequency in flanking regions in single cells, as well as a periodical pattern that supports the rotational positioning of nucleosomes[4,16] (Extended Data Fig. 8a). Smaller variation in nucleosome positioning is associated with lower frequencies of CC, GG and GC and higher frequencies of AA, TT, AT and TA in the flanking region (Extended Data Fig. 8b–e). We next explored the relationship between variance in DNA sequence and variance in nucleosome positioning. Our analysis shows that sequences occupied by nucleosomes have a higher fraction of alternative bases than those that are occupied by subnucleosome-sized particles, by tags from DNase-seq or by tags from CTCF chromatin immunoprecipitation with sequencing (ChIP-seq) (Extended Data Fig. 8f, g), which supports the notion that sequence variants influence transcription-factor binding[27] and nucleosome positioning[28]. We found that single-base variance within nucleosome regions is positively correlated with nucleosome variance across cells (Extended Data Fig. 9h). Furthermore, the single-base variance at transcription-factor motifs is positively correlated with nucleosome variance at DHSs and is also positively correlated with gene expression variation across different single cells (Extended Data Fig. 9i, j).

Enhancers display remarkable cell-type specificity. Consistent with a previous observation that active enhancers are associated with a nucleosome loss[3], the naive CD4 T cell-specific enhancers displayed decreased nucleosome occupancy in naive CD4 T cells, as revealed by the pooled scMNase-seq data from naive CD4 T cells; by contrast, enhancers that are specific to T helper 1 (T_H1) and T helper 2 (T_H2) cells showed only a very minor overall nucleosome loss in naive CD4 T cells (Extended Data Fig. 9a, b). However, examination of the nucleosome patterns at the T_H1- and T_H2-specific enhancers across different single cells revealed that 19% and 29% of naive CD4 T cells showed decreased nucleosome occupancy—which is independent of fragment-size ratio—at the de novo enhancers of T_H1 and T_H2 cells, respectively, whereas much smaller fractions of mouse ESCs and NIH3T3 cells showed decreased nucleosome occupancy at these enhancers (Fig. 4a, Extended Data Fig. 9c–e). Furthermore, subgroups of T cells that show decreased nucleosome occupancy at the T_H1 and T_H2 enhancers do not have much overlap (Extended Data Fig. 9f), which suggests they are specifically primed for the corresponding lineages. The T_H1-specific enhancers with the most nucleosome loss in naive CD4 T cells are linked to genes that encode T_H1 cytokine (*Ifng*) and key regulators (*Tbx21*, *Stat1* and *Stat4*) (Extended Data Fig. 9g, h); the T_H2-specific enhancers with the most nucleosome loss are linked to genes that encode key regulators for T_H2 differentiation (*Il4* and *Stat6*) (Extended Data Fig. 9i, j). Motif analysis revealed that the nucleosome loss at T_H1 enhancers is specifically associated with motifs for RELA, which promotes T_H1 differentiation; the nucleosome loss at T_H2 enhancers is specifically associated with motifs for GATA3 and STAT6, which promote T_H2 differentiation (Extended Data Fig. 9k). Gene Ontology analysis revealed that the higher-ranked nucleosome losses at both T_H1 and T_H2 enhancers are associated with functions in T cell differentiation, immune system process and cytokine production (Extended Data Fig. 9l, m). These results suggest that a large fraction of naive CD4 T cells have already experienced differentiating

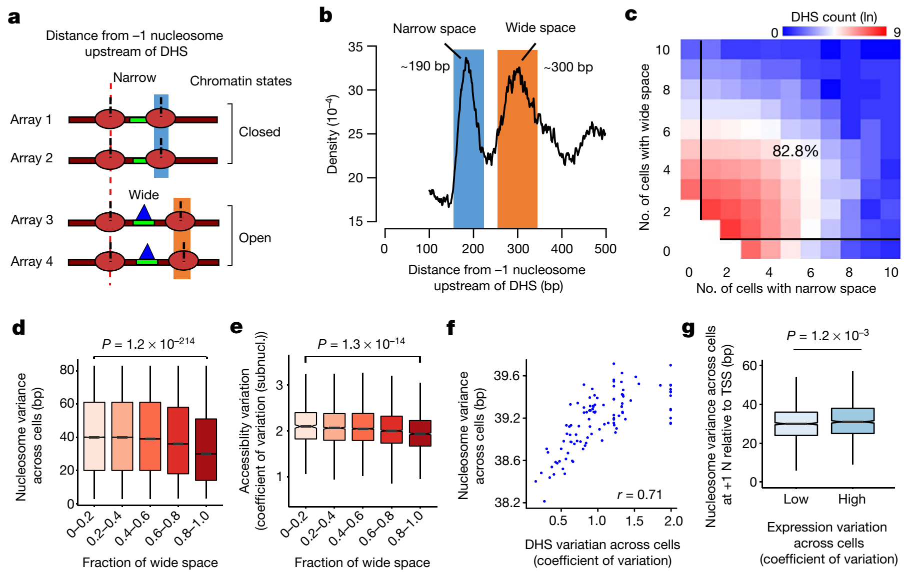


Fig. 3 | The bimodal distribution of nucleosome spacing across DHSs is associated with the cell-to-cell variation in nucleosome positioning and chromatin accessibility. a, Schema of nucleosome spacing across a DHS and two chromatin states inferred by nucleosome spacing. **b**, Density plot of nucleosome spacing across a DHS within single cells reveals two peaks that correspond to narrow spacing (blue) and wide spacing (red). **c**, Heat map showing DHS frequency as a function of number of cells with narrow spacing and number of cells with wide spacing. The percentage of DHSs in which there are both types of spacing across a DHS in different single cells is shown. **d**, **e**, Box plots showing the cell-to-cell variation in nucleosome positioning (**d**) and chromatin accessibility (**e**) for five groups of DHSs, defined by fraction of wide space. Data represent 612, 2,088, 3,858, 2,500 and 1,586 DHSs (from left to right). **f**, Scatter plot showing nucleosome variance (*y* axis) and DHS variation (*x* axis) across cells for 106 bins of DHSs, based on DHS variation. Each dot represents the average of 500 DHSs for each bin. Pearson's correlation was calculated. **g**, Box plot showing nucleosome variation at +1 nucleosome relative to TSS for two groups of genes sorted by expression variation. Low, bottom 25% ($n = 1{,}171$ genes); high, top 25% ($n = 1{,}174$ genes). In **d**, **e** and **g**, *P* values were calculated by one-sided Mann–Whitney *U*-test. In the box plots, centre line is median; boxes, first and third quartiles; whiskers, 1.5× interquartile range; notch, 95% confidence interval of the median.

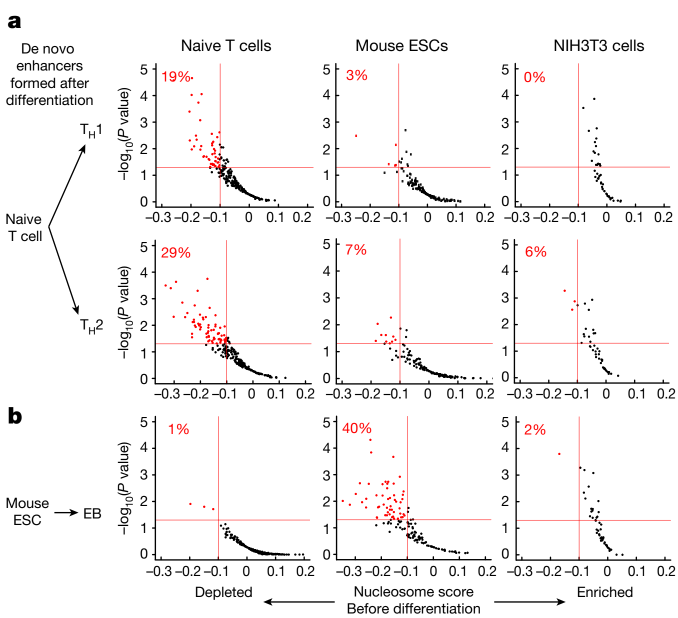


Fig. 4 | A subgroup of undifferentiated cells shows a nucleosome signature primed for differentiation. a, **b**, A large fraction of naive CD4 T cells shows decreased nucleosome occupancy at the de novo enhancers that are formed either in T_H1 (**a**, top) or T_H2 cells (**a**, bottom), whereas only a small fraction of mouse ESCs and NIH3T3 cells shows nucleosome depletion at the same enhancers. By contrast, a large fraction of mouse ESCs shows depleted nucleosomes at the de novo enhancers that are formed in EBs, whereas only a small fraction of naive CD4 T cells and NIH3T3 cells shows nucleosome depletion at the same enhancers. The fractions of primed cells are shown in red. Data represent 237 single naive T cells, 143 single mouse ESCs and 48 single NIH3T3 cells.

signalling events during the developmental history of these cells, which have primed the de novo enhancers of T_H1 or T_H2 cells by means of decreased nucleosome occupancy in the undifferentiated naive CD4 T cells.

Similarly, mouse ESCs displayed a substantial nucleosome loss at the mouse ESC-specific enhancers but only a minor loss at embryoid-body (EB)-specific enhancers, which are formed de novo after differentiation from mouse ESCs (Extended Data Fig. 10a, b). Analysis of single cells revealed that 40% of mouse ESCs showed decreased nucleosome occupancy at the de novo EB-specific enhancers, whereas only 1% and 2% of naive CD4 T cells and NIH3T3 cells, respectively, showed decreased nucleosome occupancy at these enhancers (Fig. 4b, Extended Data Fig. 10c, d). The EB enhancers with the most nucleosome loss are linked to genes that include mesoderm markers (*Brachyury* (also known as *T*) and *Wnt3*) and endoderm markers (*Gata4* and *Gata6*) (Extended Data Fig. 10e, f), and are associated with stem cell differentiation and development of various lineages, such as myeloid, neural tube and placental cells (Extended Data Fig. 10g). These results reveal the heterogeneity of cultured mouse ESCs, and suggest that some of these cells are already primed for differentiation by the reorganization of their nucleosome structure at enhancers formed in the differentiating EBs.

Here we introduce scMNase-seq, a powerful method for simultaneously measuring chromatin accessibility and nucleosome positioning in single cells, which may be paired with existing approaches—such as single-cell RNA-seq[9], single-cell DNase-seq[12] and/or single-cell ChIP-seq[29]—for systems analysis and to provide further insights into the molecular underpinning of cellular heterogeneity. Our application of scMNase-seq to three types of single cells revealed principles of nucleosome organization in different chromatin regions as well as heterogeneity of nucleosome positioning and spacing at DHSs. Our data suggest that the cellular heterogeneity of undifferentiated

cells is related to heterogeneous nucleosome organization in critical regulatory regions, which reflects the differentiation potential of these cells.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability

Custom codes for the quantification of the uniformity of nucleosome spacing and calculation of nucleosome occupancy score are available at https://github.com/binbinlai2012/scMNase.

Data availability

The scMNase-seq datasets have been deposited in the Gene Expression Omnibus database with accession number GSE96688.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0567-3.

Received: 20 March 2017; Accepted: 20 August 2018;
Published online 26 September 2018.

1. Shivaswamy, S. et al. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol*. **6**, e65 (2008).
2. Lam, F. H., Steger, D. J. & O'Shea, E. K. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246–250 (2008).
3. Schones, D. E. et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
4. Albert, I. et al. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
5. Mavrich, T. N. et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res*. **18**, 1073–1083 (2008).
6. Mavrich, T. N. et al. Nucleosome organization in the *Drosophila* genome. *Nature* **453**, 358–362 (2008).
7. Gaffney, D. J. et al. Controls of nucleosome positioning in the human genome. *PLoS Genet*. **8**, e1003036 (2012).
8. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
9. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
10. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
11. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
12. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146 (2015).
13. Kaplan, N., Hughes, T. R., Lieb, J. D., Widom, J. & Segal, E. Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biol*. **11**, 140 (2010).
14. Valouev, A. et al. Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–520 (2011).
15. Zhang, Z. et al. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **332**, 977–980 (2011).
16. Brogaard, K., Xi, L., Wang, J. P. & Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496–501 (2012).
17. Valouev, A. et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res*. **18**, 1051–1063 (2008).
18. Hughes, A. L., Jin, Y., Rando, O. J. & Struhl, K. A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. *Mol. Cell* **48**, 5–15 (2012).
19. Teif, V. B. et al. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol*. **19**, 1185–1192 (2012).
20. Szerlong, H. J. & Hansen, J. C. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochem. Cell Biol*. **89**, 24–34 (2011).
21. Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145–155 (1997).
22. Fan, Y. et al. Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **123**, 1199–1212 (2005).
23. Öberg, C., Izzo, A., Schneider, R., Wrange, Ö. & Belikov, S. Linker histone subtypes differ in their effect on nucleosomal spacing in vivo. *J. Mol. Biol*. **419**, 183–197 (2012).
24. Hartley, P. D. & Madhani, H. D. Mechanisms that specify promoter nucleosome location and identity. *Cell* **137**, 445–458 (2009).
25. Hu, G. et al. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res*. **21**, 1650–1658 (2011).
26. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
27. Maurano, M. T. et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet*. **47**, 1393–1401 (2015).
28. Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M. & Park, P. J. Impact of chromatin structure on sequence variability in the human genome. *Nat. Struct. Mol. Biol*. **18**, 510–515 (2011).
29. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol*. **33**, 1165–1172 (2015).

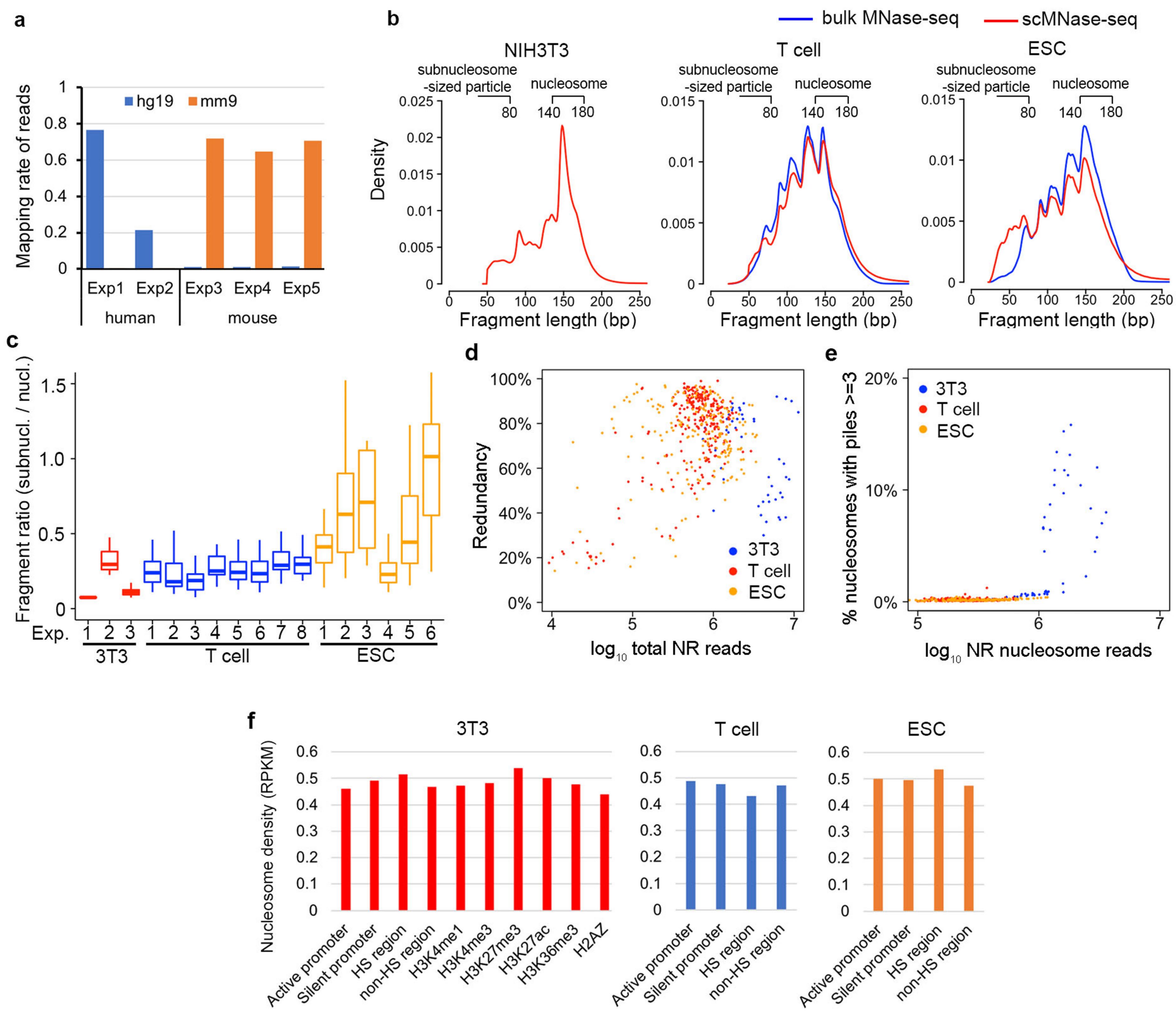
Acknowledgements We thank B. Z. Stanton for critical reading of the manuscript, B. Z. Stanton and J. Cooper for discussions, the National Heart, Lung, and Blood Institute DNA Sequencing Core Facility for sequencing the libraries, the National Heart, Lung, and Blood Institute Flow Cytometry Core facility for sorting the cells, and the National Institutes of Health Biowulf High Performance Computing Systems for computing service. The work was supported by Division of Intramural Research, National Heart, Lung and Blood Institute.

Reviewer information *Nature* thanks M. Buck and the other anonymous reviewer(s) for their contribution to the peer review of this work.

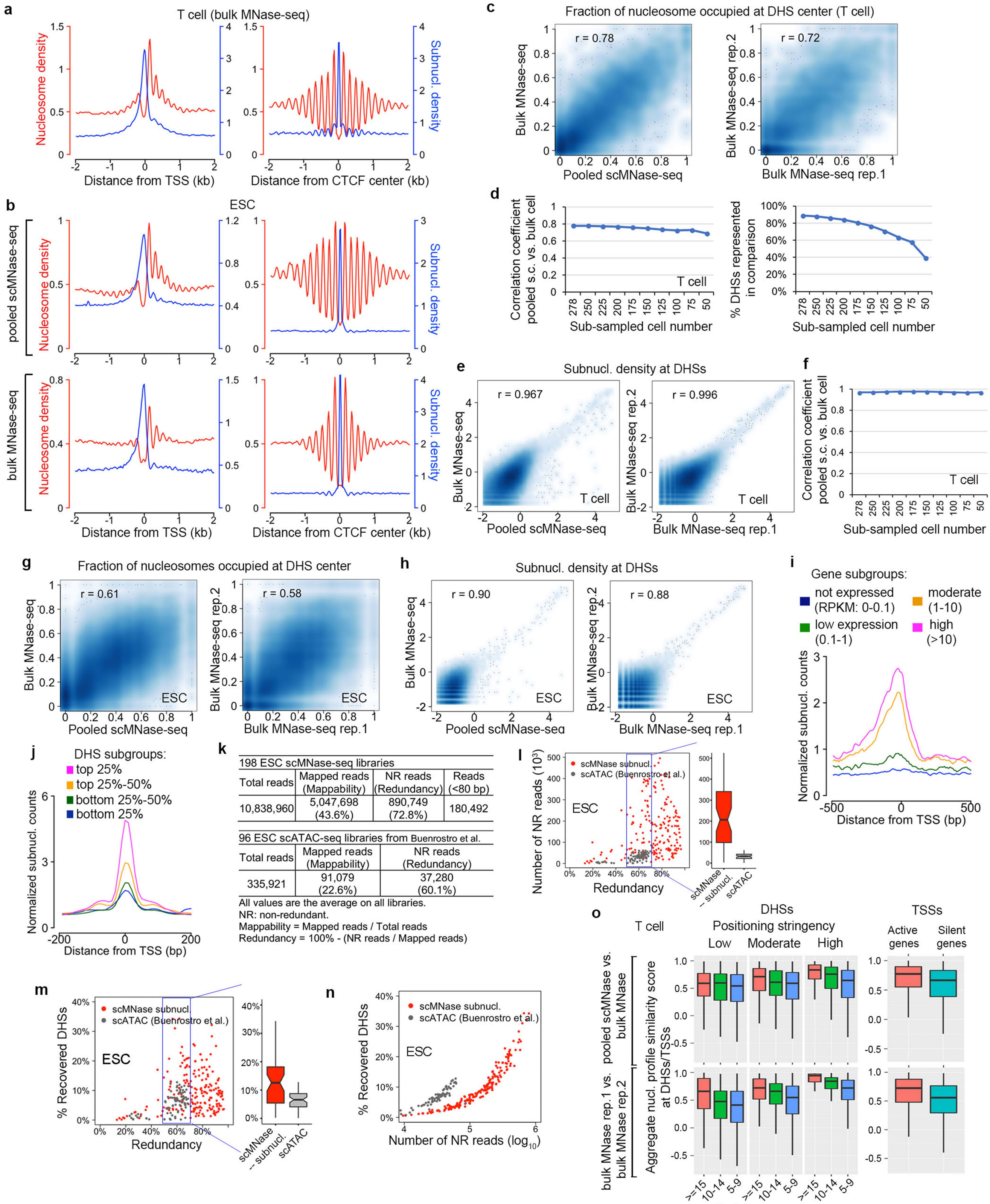
Author contributions K.Z. conceived the project. Q.T. and K.Z. performed the scMNase-seq experiments for NIH3T3 cells. K.C., W.G. and W.X. performed the scMNase-seq experiments for naive T cells and mouse ESCs. B.L. analysed the data. W.J., G.H. and B.N. contributed to the data analysis. B.L. and K.Z. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0567-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0567-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

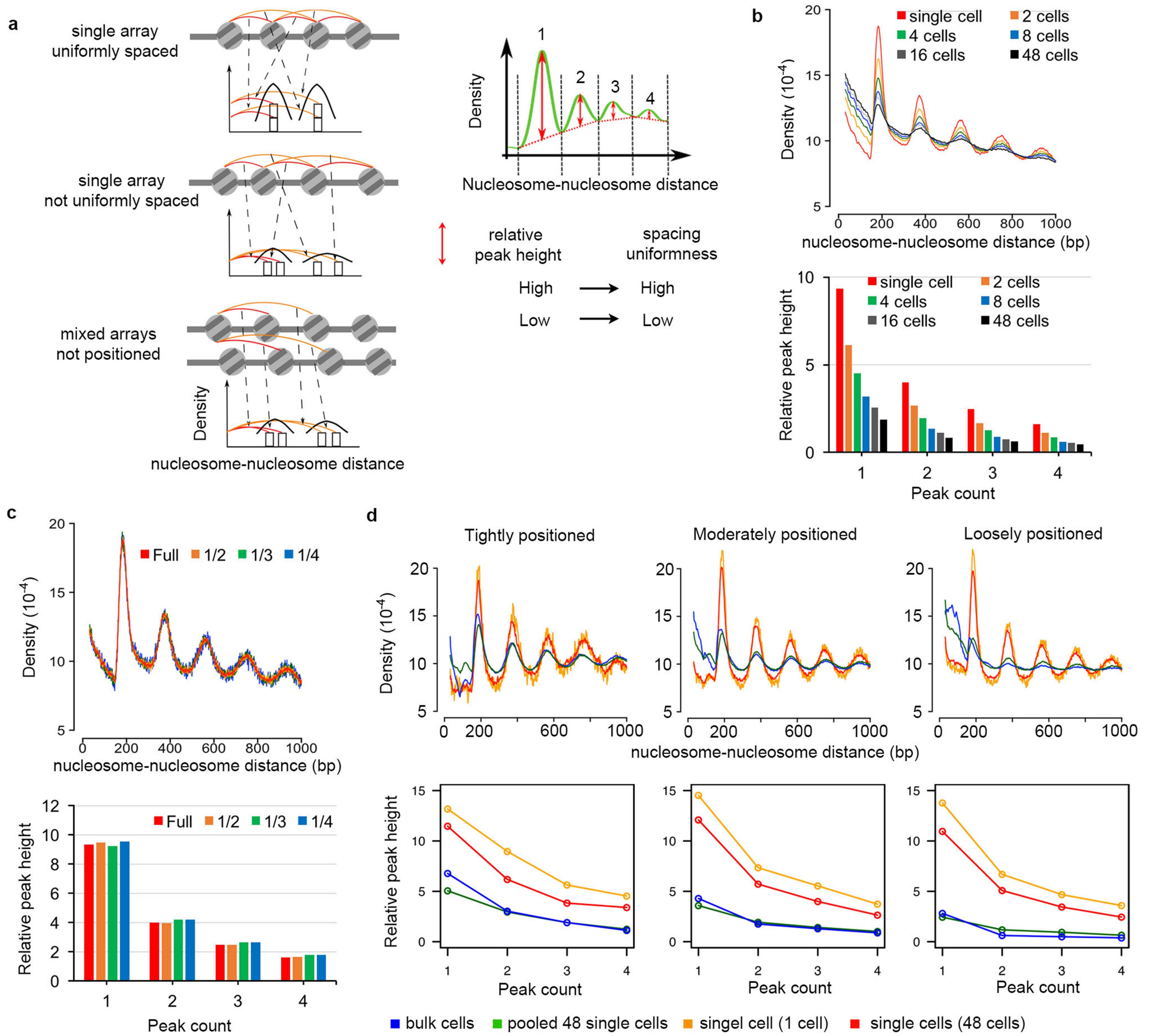


Extended Data Fig. 1 | Characterizing scMNase-seq datasets. **a**, Mapping rates of reads from 100 human cells (two experiments on the left) or 100 mouse cells (three experiments on the right) against human genome (blue) and mouse genome (orange) are shown. The cells were sorted from pre-mixed and MNase-digested human and mouse cells. These data show that there is little contamination of DNA of one cell from another cell. **b**, Fragment-length density of pooled scMNase-seq for NIH3T3 cells, pooled scMNase-seq and bulk-cell MNase-seq for T cells and mouse ESCs. **c**, Box plots of fragment ratio (subnucleosome-sized particle-to-nucleosome) for NIH3T3 cell, naive CD4 T cell and mouse ESC scMNase-seq libraries. Single-cell libraries were grouped by biologically independent experiments. Supplementary Table 1 gives the library number for each group. Centre line, median; boxes, first and third quartiles; whiskers, 1.5× interquartile range. **d**, Plot of non-redundant (NR) read number (*x* axis) and sequencing redundancy (*y* axis) for single NIH3T3 cells, CD4 T cells and mouse ESCs. **e**, Plot of non-redundant nucleosome reads (*x* axis) and percentage of nucleosomes with overlapping piles ≥ 3 (*y* axis). The plot suggests the polyploidy of NIH3T3 cells. **f**, Nucleosome density at different genomic regions for NIH3T3 cell, CD4 T cell and mouse ESC scMNase-seq libraries reveals that the nucleosomes in different genomic regions were similarly detected and represented by scMNase-seq.

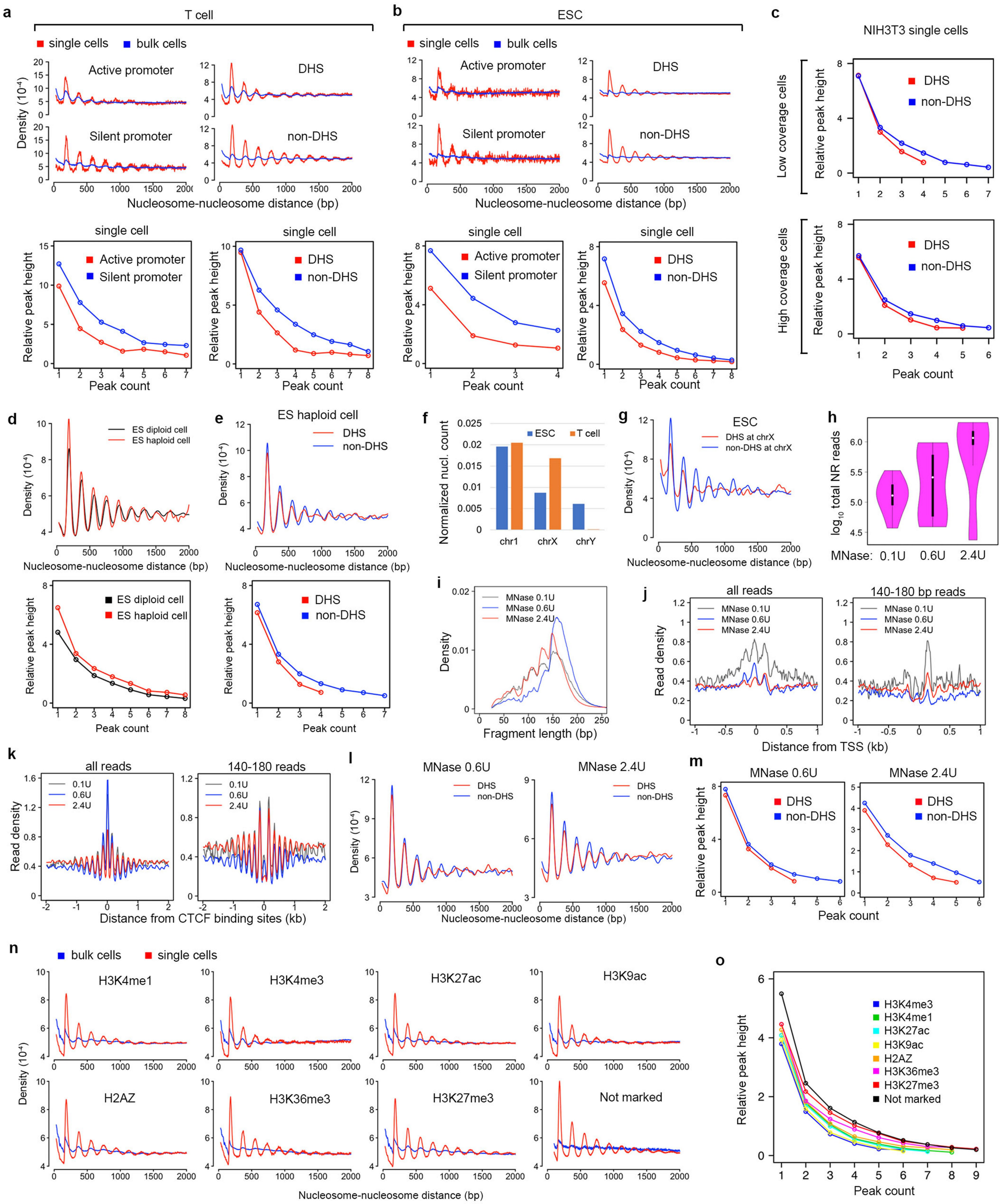


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Characterizing pooled scMNase-seq data and subnucleosome-sized particles. a, Average density profiles of nucleosomes (red) and subnucleosome-sized particles (blue) relative to the TSSs of active genes (left) and CTCF-binding sites (right) for bulk-cell naive CD4 T cell MNase-seq data. **b**, Average density profiles of nucleosomes (red) and subnucleosome-sized particles (blue) relative to the TSSs of active genes (left) and CTCF-binding sites (right) for pooled mouse ESC scMNase-seq data (top) and bulk-cell mouse ESC MNase-seq data (bottom). **c**, Smoothed scatter plot for the fraction of nucleosome occupied at 8,929 DHS centres (selected from the top 10,000 DHSs (see Supplementary Methods for criteria)) for pooled scMNase-seq (*x* axis) versus bulk-cell MNase-seq (*y* axis) for T cells (left). Pearson correlation coefficient is indicated. As a positive control, the scatter plots for two bulk-cell MNase-seq replicates are also shown (right). **d**, Pearson correlation coefficient for the fraction of nucleosomes occupied at the DHS centre between pooled sub-sampled CD4 T cell scMNase-seq libraries and bulk-cell MNase-seq, as a function of sub-sampled cell number (left). Percentages of top 10,000 DHSs represented in the comparison—that is, the sample size in the top panel—as a function of sub-sampled cell number are also shown (right). **e**, Smoothed scatter plot for subnucleosome-sized particle density at 83,229 DHSs for pooled scMNase-seq (*x* axis) versus bulk-cell MNase-seq (*y* axis) for T cells (left). Pearson correlation coefficient is indicated. As a positive control, the scatter plots for two bulk-cell MNase-seq replicates are also shown (right). **f**, Pearson correlation coefficient for subnucleosome-sized particle density between pooled sub-sampled T cell scMNase-seq libraries and bulk-cell MNase-seq at 83,229 DHSs, as a function of sub-sampled cell number. **g**, **h**, Smoothed scatter plot for the fraction of nucleosomes occupied at 8,449 DHS centres (selected from top-10,000 DHSs (see Supplementary Methods for criteria)) (**g**) and subnucleosome-sized particle density at 94,250 DHSs (**h**) for pooled scMNase-seq (*x* axis) versus bulk-cell MNase-seq (*y* axis) for mouse ESCs. Pearson correlation coefficient is indicated. As a positive control, the scatter plots for two bulk-cell MNase-seq replicates are also shown. **i**, **j**, Average density profiles of subnucleosome-sized particles around TSSs for gene subgroups with different expression levels (**i**) and around DHSs for DHS subgroups with different DNase I tag densities (**j**). **k**, Table showing the mapping statistics for 198 mouse ESC scMNase-seq libraries and 96 previously published[10] mouse ESC scATAC-seq libraries. **l**, **m**, Scatter plots of the number of non-redundant reads (**l**, *y* axis) and percentage of recovered DHSs (**m**, *y* axis) versus sequencing redundancy (*x* axis) for scMNase-seq subnucleosome-sized particles (red, $n = 198$ single-cell libraries) and scATAC-seq reads (grey, $n = 96$ single-cell libraries). Box plots (right) show the values from scatter plots (left) for cells with redundancy that ranges from 50% to 70% (blue rectangle in the left panel; red, $n = 49$; grey, $n = 58$) for the two methods. **n**, Scatter plot showing the percentage of recovered DHSs (*y* axis) versus number of non-redundant reads for scMNas-seq subnucleosome-sized particles (red, $n = 198$ single-cell libraries) and scATAC-seq reads (grey, $n = 87$ single-cell libraries). **o**, Aggregated nucleosome profile similarity score at DHSs for different groups of DHSs (left) and two promoter groups (right), for comparison between pooled scMNase-seq and bulk-cell MNase-seq (top) and between two bulk-cell MNase-seq replicates (bottom). The DHS groups are classified by three positioning-stringency levels (low, positioning score < 0.45; moderate, $0.45 <$ positioning score < 0.65, high, positioning score > 0.65) and three nucleosome coverage levels (high, ≥ 15; moderate, 10–15; low, 5–9). The DHS numbers for each group are: low positioning score and high coverage, $n = 803$; low positioning score and moderate coverage, $n = 531$; low positioning score and low coverage, $n = 450$; moderate positioning score and high coverage, $n = 701$; moderate positioning score and moderate coverage, $n = 592$; moderate positioning score and low coverage, $n = 588$; high positioning score and high coverage, $n = 162$; high positioning score and moderate coverage, $n = 230$; high positioning score and low coverage, $n = 395$. The number of promoters for each group: active, $n = 6,777$; silent, $n = 418$. In box plots in **i**, **m**, **o**, centre line, median; boxes, first and third quartiles; whiskers, 1.5× interquartile range; notch, 95% confidence interval of the median.

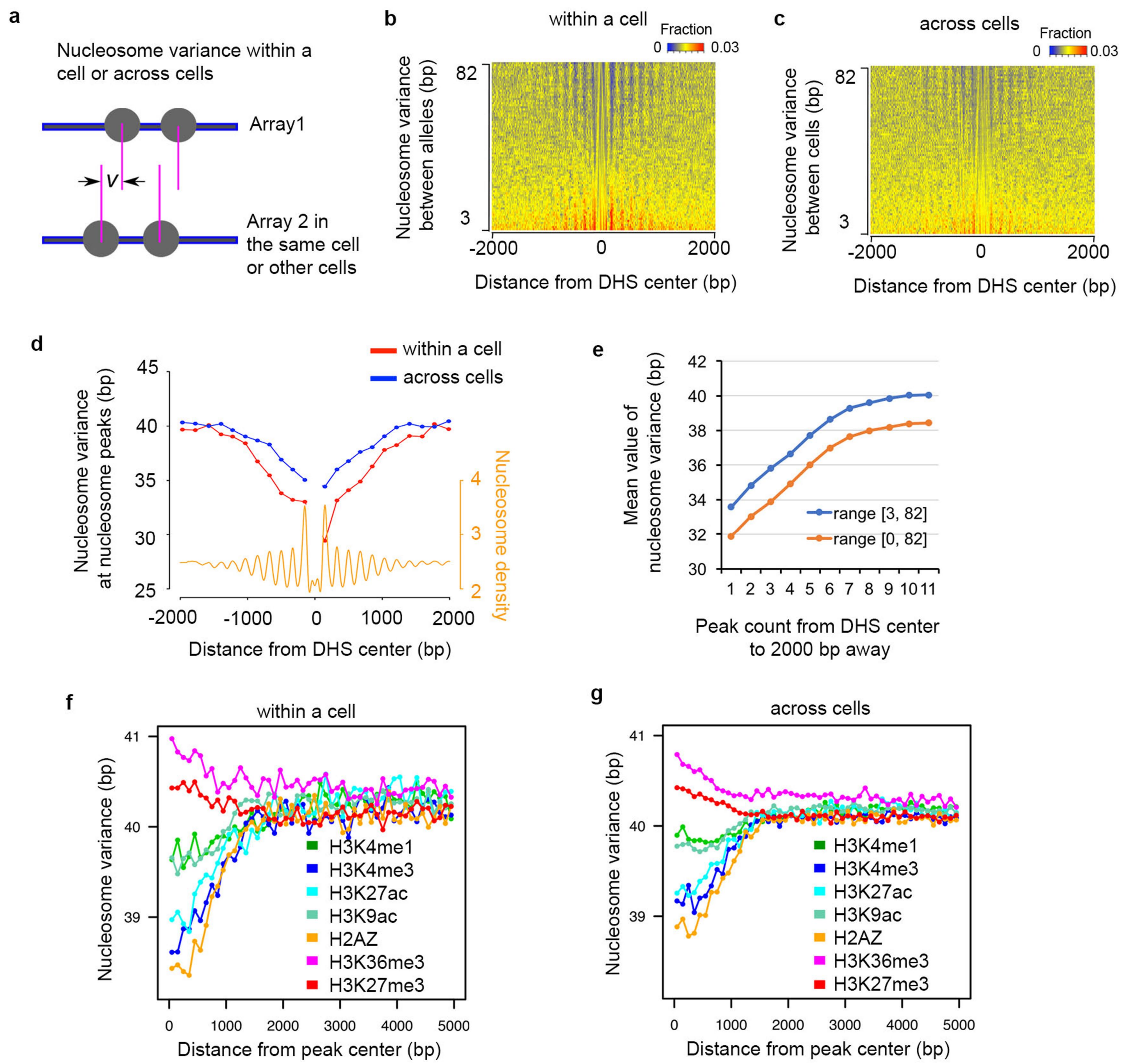


Extended Data Fig. 3 | Measuring uniformity in nucleosome spacing in single cells. a, Cartoon illustrates that uniformity in nucleosome spacing can be measured by nucleosome-to-nucleosome distance density: uniformly spaced nucleosomes in a single array result in sharp and high peaks, whereas non-uniformly spaced nucleosomes result in flat peaks or no peaks. Nucleosomes from mixed arrays also result in flat peaks, even if they are uniformly spaced. **b**, The nucleosome space phasing and relative peak height gradually decreased as the number of mixed cells increases, which indicates cellular heterogeneity of nucleosome organization across different cells. **c**, Nucleosome space phasing and relative peak height do not change when reducing the library size (number of sequence reads) to 1/2, 1/3 and 1/4. **d**, Density plots of nucleosome-to-nucleosome distance (top) and relative peak height in density plot (bottom) for nucleosomes with different positioning stringency for bulk-cell MNase-seq, pooled 48 single cells, one representative single cell and 48 single-cell scMNase-seq datasets.

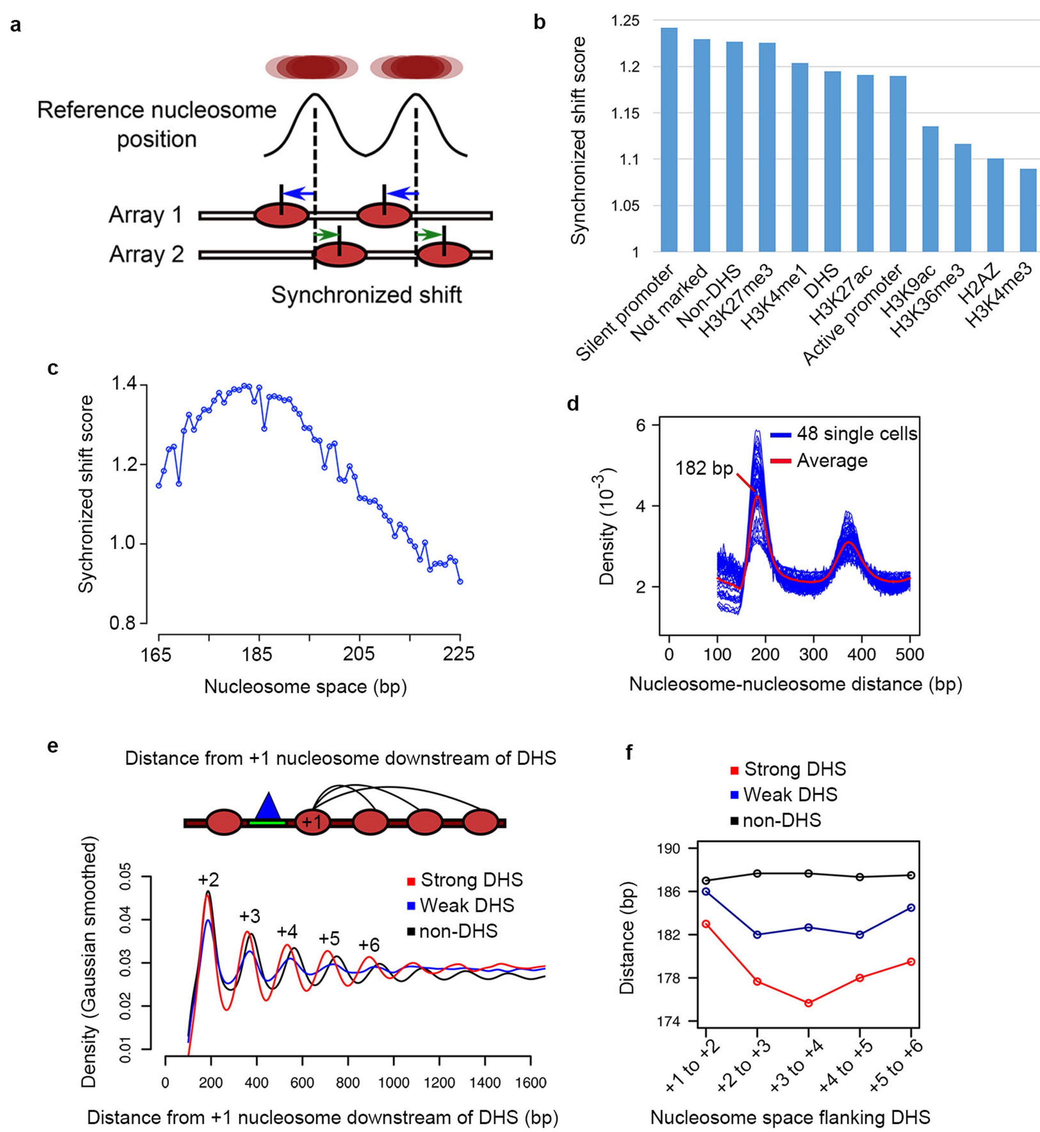


Extended Data Fig. 4 | See next page for caption.

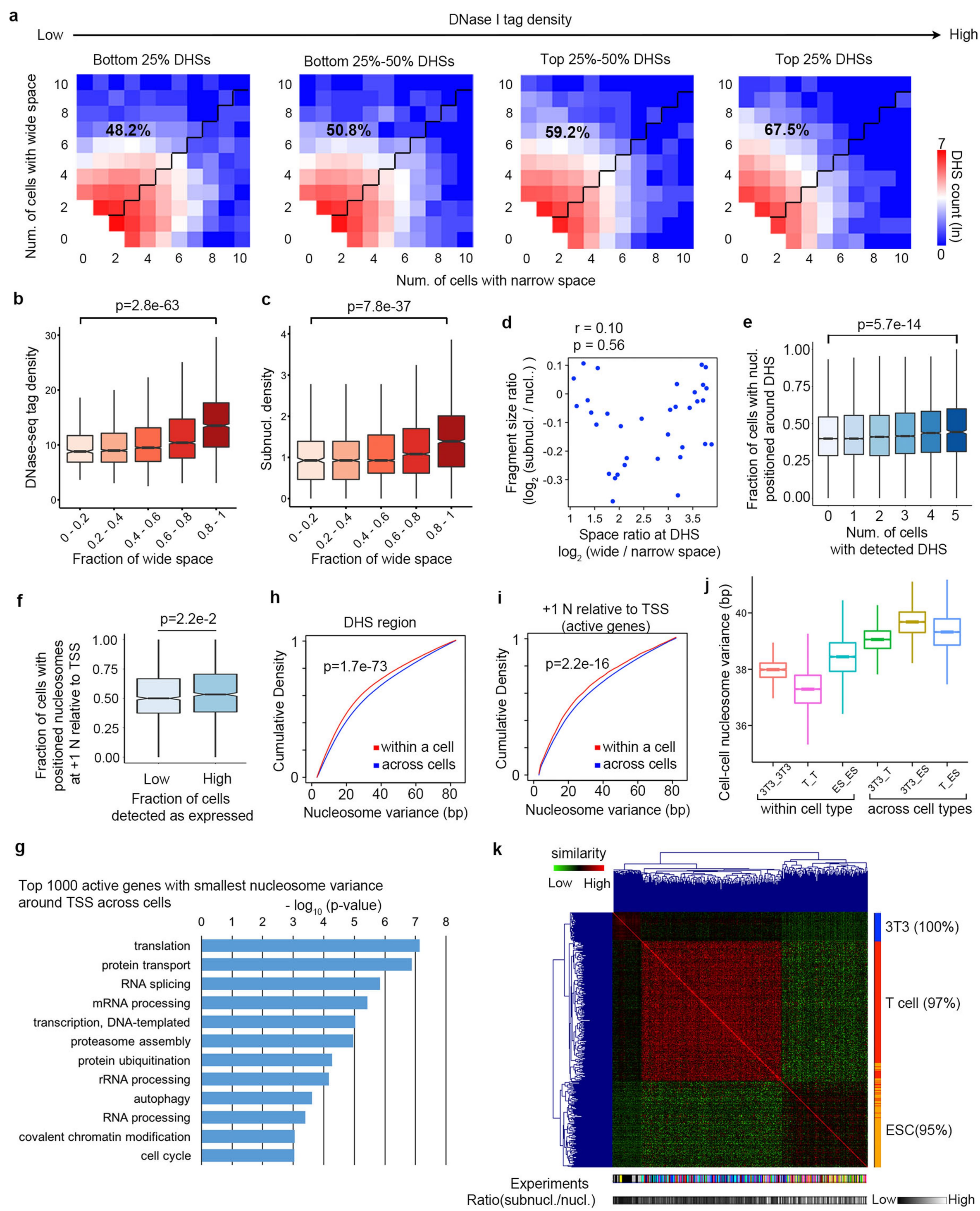
Extended Data Fig. 4 | Uniformity in nucleosome spacing is higher in silent heterochromatin regions than in active chromatin regions. **a**, **b**, Density plots of nucleosome-to-nucleosome distance (top) and relative peak height (bottom) for nucleosomes at active or silent promoters and DHS or non-DHS regions for T cells (**a**) and mouse ESCs (**b**). **c**, Relative peak height of density plots for nucleosome-to-nucleosome distance for nucleosomes in DHS (red) and non-DHS regions (blue) for low-coverage cells (top) and high-coverage cells (bottom). **d**, Density plots of nucleosome-to-nucleosome distance (top) and relative peak height (bottom) for diploid (black) and haploid (red) mouse ESCs. **e**, Density plots of nucleosome-to-nucleosome distance (top) and relative peak height (bottom) at DHS and non-DHS regions for haploid mouse ESCs. **f**, Mapped nucleosome count normalized by chromosome length at chromosome 1, X and Y for mouse ESCs and CD4 T cells suggests that mouse ESCs are derived from male mouse. **g**, Density plots of nucleosome-to-nucleosome distance at DHS and non-DHS regions at chromosome X for mouse ESCs. **h**, Violin plots of library size (total non-redundant reads) for NIH3T3 scMNase-seq libraries treated with three MNase concentrations (0.1 unit (0.1 U), 0.6 unit (0.6 U) and 2.4 unit (2.4 U) MNase per million cells). Each condition has 10 single-cell libraries. In the violin plots, centre dot, mean; inner layer, the interquartile range. **i**, Fragment-length density of pooled scMNase-seq data with three MNase concentrations. **j**, **k**, Average density profiles of all reads (**j**, **k**, left) and nucleosome reads with length between 140 and 180 bp (**j**, **k**, right) around the TSSs of active genes (**j**) and CTCF-binding sites (**k**) for pooled scMNase-seq with three MNase concentrations. **l**, **m**, Density plots of nucleosome-to-nucleosome distance (**l**) and relative peak height (**m**) at DHS (red) and non-DHS (blue) regions for scMNase-seq treated by 0.6 U (left) and 2.4 U (right) MNase concentrations. **n**, **o**, Density plots of nucleosome-to-nucleosome distance (**n**) and relative peak height (**o**) for nucleosomes within genomic regions marked by different histone modifications.



Extended Data Fig. 5 | Variation in nucleosome positioning within a cell or across different single cells around the centre of DHS or histone modification peaks. **a**, Cartoon illustrating the definition of nucleosome variance within a cell or across different single cells. **b**, **c**, Heat maps showing the distribution of nucleosome variance at the position relative to DHS centre within a cell (**b**) or across different single cells (**c**). **d**, Nucleosome variance within a cell (red) and across single cells (blue) becomes smaller when getting closer to the DHS centre. **e**, Calculations of mean value of nucleosome variance from two ranges (3–82 bp and 0–82 bp) reveal the same trend of increase when nucleosomes become farther away from DHS centre. **f**, **g**, Average profiles of nucleosome variance at the position relative to the centre of histone modification peaks within a cell (**f**) or across different cells (**g**).

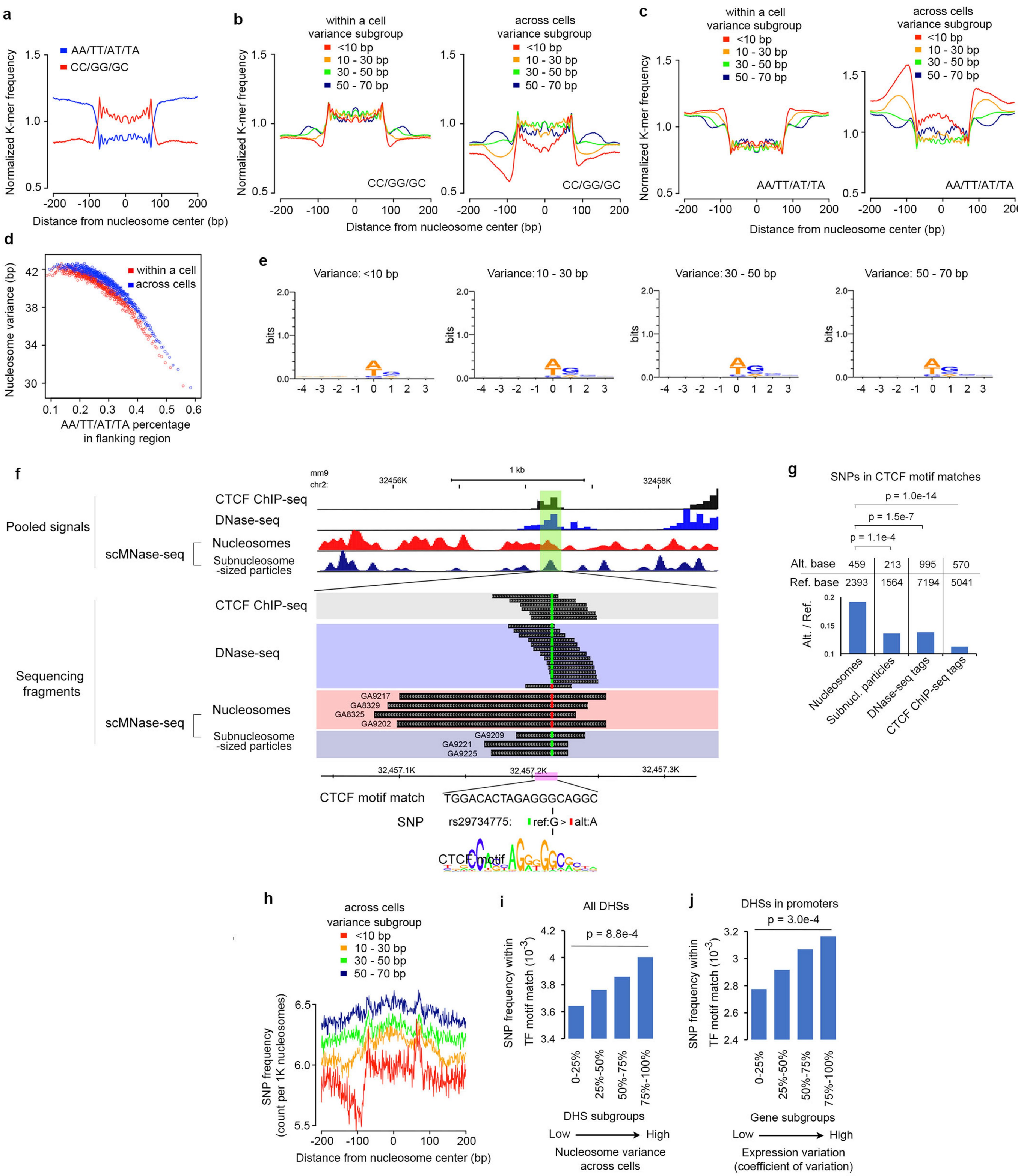


Extended Data Fig. 6 | Nucleosomes show a synchronized shift in silent-gene promoters and heterochromatin regions, and show compressed spacing where they flank DHS centres. a, Cartoon illustrates synchronized shift of adjacent nucleosomes within single nucleosome arrays. **b**, Bar plot showing synchronized shift score for different genomic regions. Silent promoter, silent-gene promoter; active promoter, active-gene promoter; not marked, regions not marked by any histone modifications as shown; DHS: ±2,000-bp region surrounding DHS centre; non-DHS, intervals of DHS regions. **c**, Synchronized shift score for nucleosome pairs with different distances of nucleosome space. **d**, Density plot of nucleosome-to-nucleosome distance in single cells reveals dominant nucleosome space at about 182 bp. **e**, Density plot of nucleosome spacing in the regions flanking strong and weak DHSs as well as non-DHSs. **f**, Distances between each pair of nucleosomes in the chromatin regions flanking strong DHS, weak DHS or non-DHS, described in **e**.



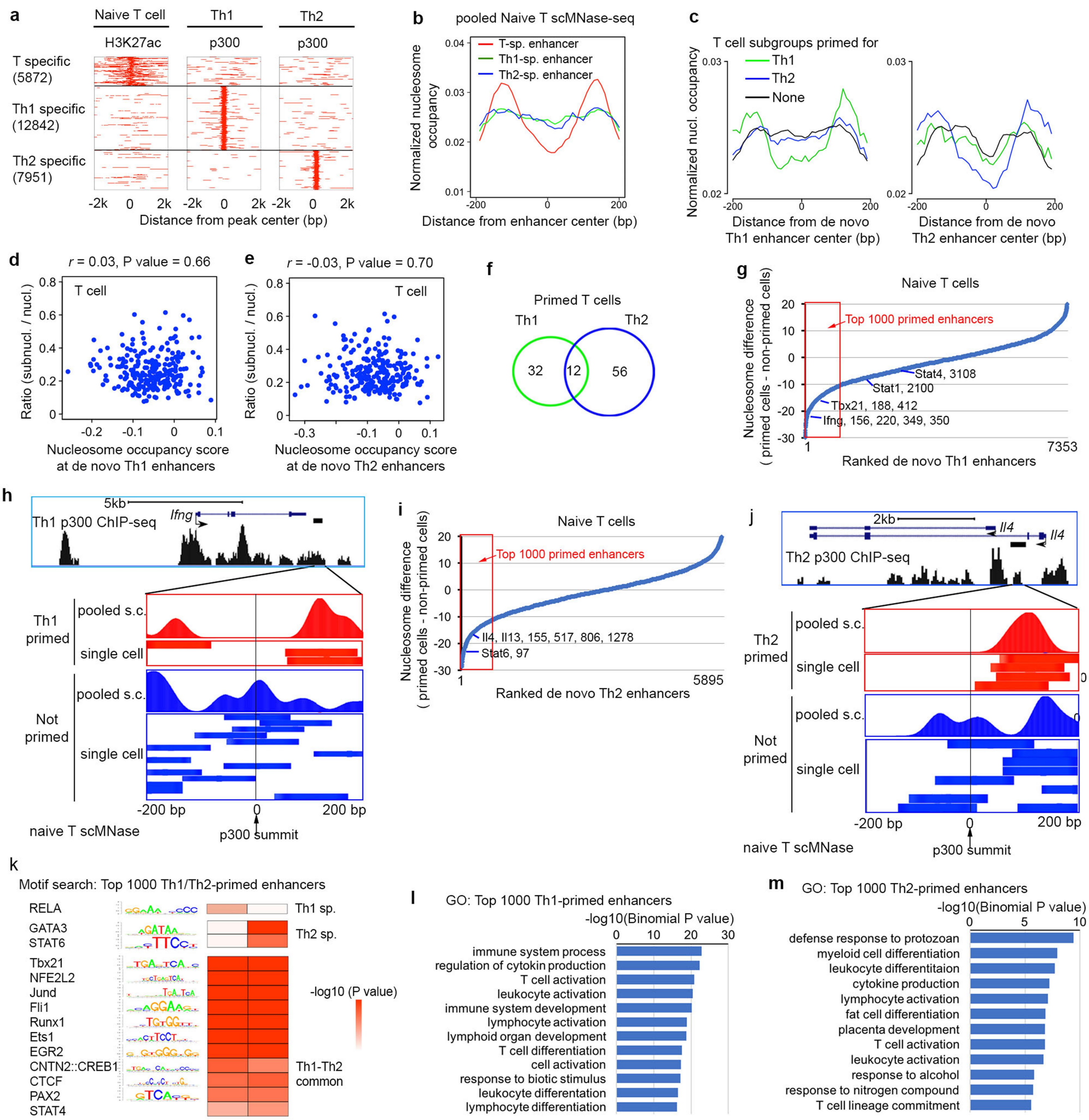
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Heterogeneity of nucleosome spacing and positioning around DHS across different single cells. a, Heat maps showing DHS frequency as a function of number of cells with the narrow spacing (*x* axis) and number of cells with the wide spacing (*y* axis) for four DHS subgroups with different DNase I tag densities. Numbers indicate the percentages of DHSs that have more wide space than narrow space. **b**, **c**, Box plots showing the accessibility level from cell population, measured by DNase-seq tag density (**b**) and pooled scMNase-seq subnucleosome-sized particle density (**c**), for five groups of DHSs defined by fraction of wide space. Data represent values on 612, 2,088, 3,858, 2,500 and 1,586 DHSs (from left to right). **d**, Scatter plot of the ratio of wide-to-narrow space at the DHS in a single cell (*x* axis) and fragment-size ratio of subnucleosome-sized particles to nucleosomes (*y* axis) on 48 NIH3T3 scMNase-seq libraries. Pearson correlation coefficient and *P* value are indicated. *P* value is the probability that one would have found the current result if the correlation coefficient were zero (null hypothesis), and was calculated using R. **e**, Box plot showing fraction of cells with positioned nucleosomes around a DHS for different groups of DHSs. DHSs were grouped on the basis of the number of cells detected as DHS in a previously published scDNase-seq experiment[12]. Number of DHSs for each group was 44,040, 15,622, 11,056, 8,009, 4,063 and 1,180 (from left to right). **f**, Box plot showing fraction of cells with a positioned +1 nucleosome for two groups of genes sorted by expression variation (low, $n = 1{,}171$; high, $n = 1{,}174$). **g**, Gene Ontology analysis of top-1,000 active genes with the smallest nucleosome variance across cells. Significant Gene Ontology terms with *P* value are reported by David Bioinformatics Resources (v.6.7). **h**, Density plot showing nucleosome variance around DHSs within a cell ($n = 73{,}274$ nucleosome pairs) and across different cells ($n = 752{,}398$ nucleosome pairs). **i**, Cumulative density plot for nucleosome variation at +1 nucleosome relative to the TSSs of active genes within a cell (red, $n = 11{,}388$ nucleosome pairs) and across cells (blue, $n = 237{,}006$). **j**, Box plot showing nucleosome variance around DHSs across cells for within a cell type (NIH3T3–NIH3T3 cells, $n = 1{,}128$ nucleosome pairs; T cells–T cells, $n = 23{,}936$; ESCs–ESCs, $n = 5{,}775$) and across different cell types (NIH3T3–T cells, $n = 11{,}856$; NIH3T3–ESCs, $n = 6{,}962$; T cells–ESCs, $n = 20{,}442$). **k**, Heat map reveals clustering of NIH3T3 cells, T cells and mouse ESCs based on cell-to-cell nucleosome dissimilarity score around DHSs. Colour bar on the right indicates cell types and colour bars on the bottom indicate experiment time and fragment-size ratio. *P* values in panels **b**, **c**, **e**, **f**, **h** and **i** were calculated using one-sided Mann–Whitney *U*-test. In box plots, centre line, median; boxes, first and third quartiles; whiskers, 1.5× interquartile range; notch, 95% confidence interval of the median.

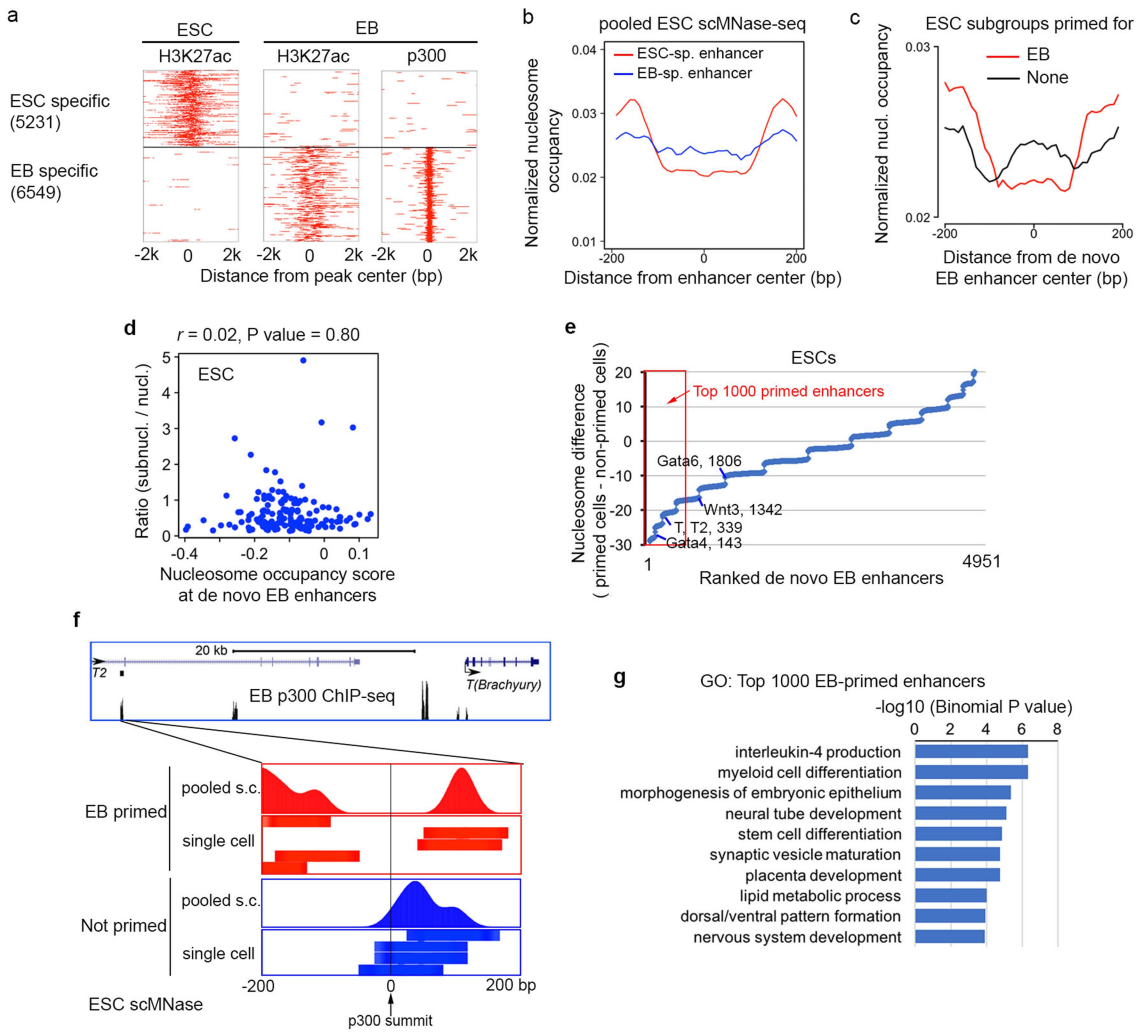


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Cell-to-cell single-base variation is associated with variation in nucleosome positioning and variation in gene expression across different single cells. **a**, CC, GG and GC frequency is higher in the nucleosome-occupied region than in the flanking region, whereas AA, TT, AT and TA frequency shows the opposite pattern. **b**, CC, GG and GC frequency in flanking regions increases as nucleosome variance within a cell (left) or across different single cells (right) increases. **c**, AA, TT, AT and TA frequency in flanking regions decreases as nucleosome variance within a cell (left) or across different single cells (right) increases. **d**, Nucleosome variances within a cell and across different single cells are reversely correlated with the percentage of AA, TT, AT and TA in flanking regions. **e**, Weblogos sequences logos for sequence preferences across MNase cleavage sites are shown for subgroups of nucleosomes with different positioning variance across cells. **f**, An example showing a CTCF motif with the reference base (green) in some cells and alternative base (red) in other cells. scMNase-seq data show that the reference base is associated with subnucleosome-sized particles, whereas the alternative base is associated with the nucleosome structure. Fragments from DNase-seq and CTCF ChIP-seq datasets within the window are also shown with the bases at single-nucleotide polymorphism location highlighted. Tracks for tag densities of CTCF ChIP-seq, DNase-seq, and nucleosomes and subnucleosome-sized particles from pooled single cells are shown in a zoomed-out window. **g**, The number of CTCF-motif matches containing alternative or reference bases at the single-nucleotide polymorphism locus occupied by nucleosomes, subnucleosome-sized particles, sequence reads obtained by DNase-seq and by CTCF ChIP-seq. *P* value was calculated using one-sided Fisher's exact test. The ratio between alternative and reference bases is also shown (bottom). **h**, Single-nucleotide polymorphism frequency is correlated with nucleosome variation across different single cells. Variant frequencies at each position relative to nucleosome midpoint for four nucleosome subgroups with different levels of nucleosome variance across cells are shown. **i**, Single-nucleotide polymorphism frequency within transcription-factor motifs at DHSs for four DHS subgroups, sorted by nucleosome variance around DHS across different single cells (each subgroup has 22,139 DHSs that contains at least one transcription-factor motif match). **j**, Single-nucleotide polymorphism frequency within transcription-factor motifs at DHSs in promoters for gene subgroups, sorted by expression variation across different single cells (each subgroup has 2,136 genes). *P* value in **i**, **j** is defined as the probability of observing a larger difference than current result between two groups by random. *P* value calculation is described in Supplementary Methods. SNP, single-nucleotide polymorphism.



Extended Data Fig. 9 | Characterization of primed enhancers in undifferentiated naive CD4 T cells. a, Heat maps show H3K27ac in naive T cells and p300 in T_H1 and T_H2 cells around naive T cell-specific, T_H1-specific and T_H2-specific enhancers. **b**, Profile of nucleosome occupancy from pooled naive T cell scMNase-seq around T cell-specific, T_H1-specific and T_H2-specific enhancers. **c**, Normalized nucleosome occupancy within ± 200 bp of the centre of de novo T_H1 enhancers (left) or de novo T_H2 enhancers (right) for subgroups of T cells primed for T_H1 cells (green), T_H2 cells (blue) or none (black). **d**, **e**, Plots of fragment-size ratio of subnucleosome-sized particles-to-nucleosomes versus nucleosome occupancy score at de novo T_H1 (**d**) and T_H2 (**e**) enhancers for 237 naive CD4 T cells reveal that nucleosome occupancy score is not correlated with fragment-size ratio. Pearson correlation coefficient and *P* value are indicated. *P* value is the probability that one would have found the current result if the correlation coefficient were zero (null hypothesis), and was calculated using R. **f**, Subgroups of naive CD4 T cells primed for T_H1 and T_H2 do not have much overlap. **g**, Plots of de novo T_H1 enhancers ranked on the basis of differences in nucleosome occupancy between pooled primed cells and the non-primed cells (*y* axis, see Supplementary Methods). Enhancers associated with key genes for T_H1 were labelled by genes along with ranks. **h**, Nucleosome positions in pooled or single primed (red) and non-primed (blue) cells at de novo T_H1-specific enhancers for *Ifng* gene. **i**, Plots of de novo T_H2 enhancers ranked on the basis of differences in nucleosome occupancy between pooled primed cells and the non-primed cells (*y* axis, see Supplementary Methods). Enhancers associated with key genes for T_H2 were labelled by genes along with ranks. **j**, Nucleosome positions in pooled or single primed (red) and non-primed (blue) cells at de novo T_H2-specific enhancers for *Il4* gene. **k**, Motifs enriched in the top 1,000 T_H1/T_H2-primed enhancers are shown. **l**, **m**, Gene Ontology analysis for top 1000 T_H1-primed (**l**) and T_H2-primed (**m**) enhancers. Significant Gene Ontology terms with P values are reported using GREAT v.3.0.0.



Extended Data Fig. 10 | Characterization of primed enhancers in undifferentiated mouse ESCs. a, Heat maps show H3K27ac in mouse ESCs and EB cells and p300 in EB cells around ESC-specific and EB-specific enhancers. **b**, Profile of nucleosome occupancy from pooled mouse ESC scMNase-seq around mouse ESC-specific and EB-specific enhancers. **c**, Normalized nucleosome occupancy within ± 200 bp of the centre of de novo EB-specific enhancers for subgroups of mouse ESCs that are primed for EB (red) or not primed for EB (black). **d**, Plots of fragment-size ratio of subnucleosome-sized particles-to-nucleosomes versus nucleosome occupancy score at de novo EB-specific enhancers for 144 mouse ESCs. Pearson correlation coefficient and *P* value are indicated. *P* value is the probability that one would have found the current result if the correlation coefficient were zero (null hypothesis), and was calculated using R. **e**, Plots of de novo EB-specific enhancers ranked on the basis of difference in nucleosome occupancy between pooled primed cells and the non-primed cells. Enhancers associated with key genes for EB cells were labelled by genes along with ranks. **f,** Nucleosome positions in pooled or single primed (red) and non-primed (blue) cells at de novo EB-specific enhancers for *Brachyury* gene. **g**, Gene Ontology analysis for top-1,000 EB-primed enhancers. Significant Gene Ontology (GO) terms with *P* values are reported using GREAT v.3.0.0.

LETTER

https://doi.org/10.1038/s41586-018-0568-2

Crystal structure of a membrane-bound *O*-acyltransferase

Dan Ma[1], Zhizhi Wang[1], Christopher N. Merrikh[2], Kevin S. Lang[2], Peilong Lu[3], Xin Li[1,4], Houra Merrikh[2,5], Zihe Rao[4,6,7,8,9] & Wenqing Xu[1,6]*

Membrane-bound *O*-acyltransferases (MBOATs) are a superfamily of integral transmembrane enzymes that are found in all kingdoms of life[1]. In bacteria, MBOATs modify protective cell-surface polymers. In vertebrates, some MBOAT enzymes—such as acyl-coenzyme A:cholesterol acyltransferase and diacylglycerol acyltransferase 1—are responsible for lipid biosynthesis or phospholipid remodelling[2,3]. Other MBOATs, including porcupine, hedgehog acyltransferase and ghrelin acyltransferase, catalyse essential lipid modifications of secreted proteins such as Wnt, hedgehog and ghrelin, respectively[4–10]. Although many MBOAT proteins are important drug targets, little is known about their molecular architecture and functional mechanisms. Here we present crystal structures of DltB, an MBOAT responsible for the D-alanylation of cell-wall teichoic acid in Gram-positive bacteria[11–16], both alone and in complex with the D-alanyl donor protein DltC. DltB contains a ring of 11 peripheral transmembrane helices, which shield a highly conserved extracellular structural funnel extending into the middle of the lipid bilayer. The conserved catalytic histidine residue is located at the bottom of this funnel and is connected to the intracellular DltC through a narrow tunnel. Mutation of either the catalytic histidine or the DltC-binding site of DltB abolishes the D-alanylation of lipoteichoic acid and sensitizes the Gram-positive bacterium *Bacillus subtilis* to cell-wall stress, which suggests cross-membrane catalysis involving the tunnel. Structure-guided sequence comparison among DltB and vertebrate MBOATs reveals a conserved structural core and suggests that MBOATs from different organisms have similar catalytic mechanisms. Our structures provide a template for understanding structure–function relationships in MBOATs and for developing therapeutic MBOAT inhibitors.

The MBOAT superfamily comprises more than 7,000 proteins (see http://pfam.xfam.org/family/MBOAT). These proteins perform divergent functions with distinct substrate preferences, although many use acyl-coenzyme A (acyl-CoA) as the acyl-group donor (Extended Data Fig. 1). Among bacterial MBOATs, DltB is essential for the D-alanylation of cell-wall teichoic acids[11–16], which are important for the growth, biofilm formation, adhesion and virulence of Gram-positive bacterial pathogens. To understand the molecular mechanisms of MBOAT proteins, we have determined the crystal structure of full-length DltB from *Streptococcus thermophilus* at 3.3 Å resolution (Fig. 1, Extended Data Figs. 2, 3, Extended Data Table 1). DltB contains 415 residues arranged into 17 helices, and both the N and the C termini are located in the extracellular space (Fig. 1a). The helices are located mostly within the lipid bilayer, with the exception of the short N- and C-terminal helices. Among them, 11 transmembrane helices form an external ring-shaped ridge, and shield a central basin that is thinner than the lipid bilayer (Fig. 1, Extended Data Fig. 4). The thin central area results from an intracellular concave surface and a more pronounced extracellular structural funnel (Fig. 1d). Because they are more conserved than the peripheral-ring helices among MBOAT proteins and are probably involved in catalysis (see below), we refer to the structural components in this thin central area as the MBOAT central core. The 3D structure of DltB can be approximately divided into three parts: the N-terminal helical ridge (N-ridge), the central core and the C-terminal helical ridge (C-ridge) (Extended Data Fig. 4). A Dali search using our DltB structure did not find any protein with a similar fold.

The extracellular side of DltB forms a structural funnel, which extends into the middle of the lipid bilayer (Fig. 1d). The surface inside the funnel is formed by residues from several transmembrane helices and loops. Notably, in sharp contrast to the low conservation of residues forming the outer-ridge surfaces, these inner residues are highly conserved among DltB proteins (Fig. 1e, Extended Data Fig. 5). Previous studies have shown that a histidine residue strictly conserved in all confirmed MBOAT proteins is probably involved in catalysis. Mutation of the corresponding histidine residue in all tested MBOATs—porcupine (PORCN), hedgehog acyltransferase (HHAT), ghrelin *O*-acyltransferase (GOAT), diacylglycerol acyltransferase 1 (DGAT1) and acyl-coenzyme A:cholesterol acyltransferase (ACAT)—either abolished or substantially reduced the acyltransferase activities of the enzymes[17–22]. In our DltB structure, this histidine residue (His336, the last residue of helix H14) is located at the bottom of the extracellular funnel (Fig. 1d, f). Another highly conserved histidine residue (His289) in the MBOAT superfamily[1] is also located at the bottom of this funnel and is spatially close to His336 (Extended Data Fig. 3). Our crystal structure and the structural conservation strongly suggest that this extracellular funnel is important for the activity of DltB.

Four *Staphylococcus aureus* DltB mutations—corresponding to *S. thermophilus* DltB mutants S165T, A209D, F250L and F250I—have been identified as resistant to the DltB inhibitors *m*-AMSA (amsacrine) and *o*-AMSA[14]. Ser165, Ala209 and Phe250 are spatially located at the surface of the funnel, with Ser165 and Phe250 sitting near the bottom of the funnel and close to His336 (Fig. 1f). We predict that *m*-AMSA and *o*-AMSA bind in this DltB funnel, and that the abovementioned four mutations may abolish inhibitor binding. We speculate that this funnel may be involved in extracellular teichoic acid substrate binding or have other key roles in catalysis. Given the biological importance of DltB[14] and the marked conservation of the extracellular funnel surface of DltB, inhibitors of DltB that bind to this funnel may act as wide-spectrum antibiotics against Gram-positive bacteria.

In addition to its role in D-alanylation, DltB also has a role in host–pathogen interactions. A missense mutation (T113K) in *S. aureus* DltB is sufficient to convert an *S. aureus* strain from a human-specific pathogen to a rabbit-specific pathogen, without any change in the D-alanylation level of lipoteichoic acid (LTA)[23]. Notably, Thr113—as well as all ten other *S. aureus* DltB residues that are associated with a change in host specificity—is located at a non-conserved extracellular apex (Extended

[1]Department of Biological Structure, University of Washington, Seattle, WA, USA. [2]Department of Microbiology, University of Washington, Seattle, WA, USA. [3]Department of Biochemistry, University of Washington, Seattle, WA, USA. [4]College of Life Sciences, State Key Laboratory of Medicinal Chemistry Laboratory, Nankai University, Tianjin, China. [5]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [6]National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. [7]Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai, China. [8]University of Chinese Academy of Sciences, Beijing, China. [9]Laboratory of Structural Biology, School of Medicine, Tsinghua University, Beijing, China. *e-mail: wxu@uw.edu

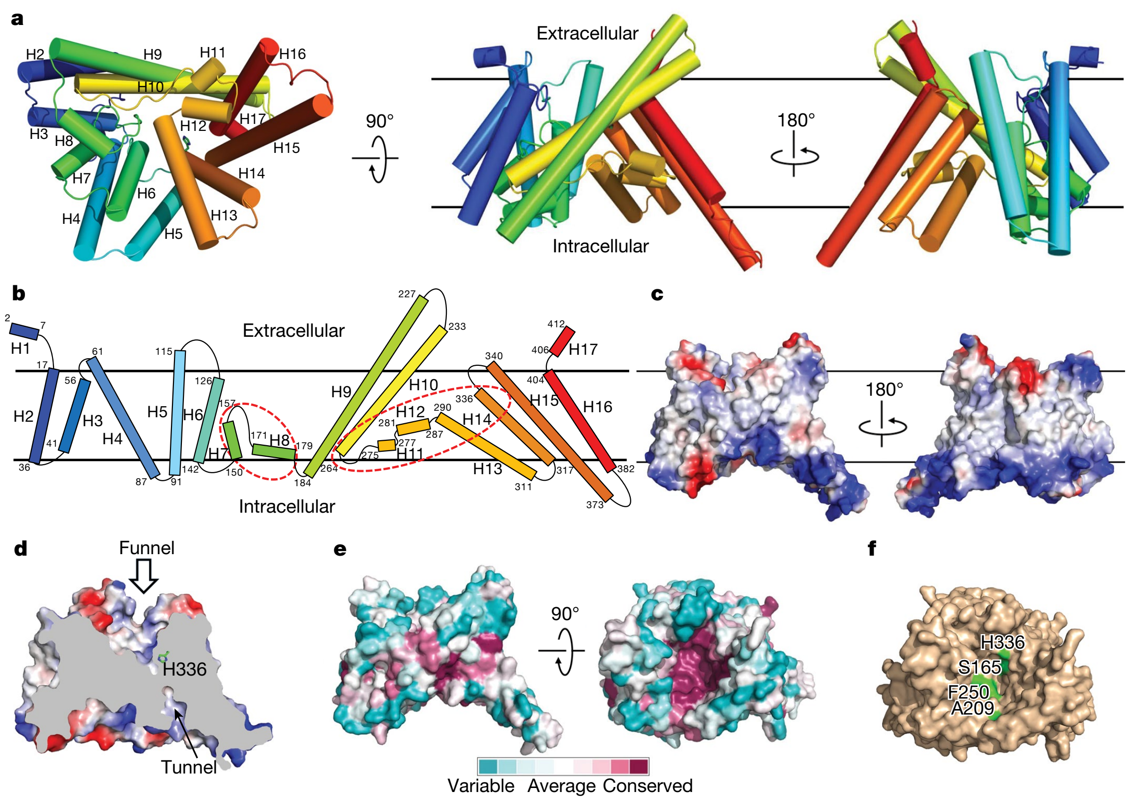


Fig. 1 | Overall structure of DltB and its conserved extracellular funnel. **a**, The DltB crystal structure is shown in three orientations with rainbow colours: bottom, front and back (from left to right). **b**, Cartoon of the transmembrane topology of DltB. DltB contains a ring of 11 peripheral transmembrane helices, which shield a central thin layer (the structural core) highlighted by two red dashed circles. **c**, The electrostatic surface of DltB. **d**, A cut-away surface illustration showing the outward funnel connected with the cytosolic side through a tunnel. The histidine residue that is completely conserved among MBOATs (His336) is located at the bottom of the funnel. **e**, Conservation of the extracellular DltB funnel surface. The surface conservation pattern was generated on the basis of sequence alignment shown in Extended Data Fig. 5. **f**, Top view of DltB showing the location of His336 and the other three DltB residues (Ser165, Ala209 and Phe250) that, when altered, were found to desensitize *S. aureus* to the inhibition of LTA D-alanylation by *m*-AMSA.

Data Figs. 4d, 5). This unusual feature strongly suggests that DltB from *S. aureus* and potentially some other species may interact with one or more unknown host factors using their extracellular ridges.

To serve as the D-alanyl donor to teichoic acid in the Dlt-mediated D-alanylation system, DltC first needs to be modified with the 4′-phosphopantetheine (Ppant) group at Ser35, a modification that can be catalysed by acyl carrier protein synthase (AcpS). The Ppant-modified DltC can be further modified with a D-alanyl group by DltA, through a thioester bond (Fig. 2a). To test whether DltB can directly interact with DltC, we co-expressed His-tagged AcpS and GST-tagged DltC. The purified DltC was uniformly modified by Ppant, as confirmed by mass spectrometry (Extended Data Fig. 2b, c). GST pull-down and size-exclusion chromatography experiments showed that DltC-Ppant and DltB form a tight complex (Fig. 2b). Octet binding analysis showed a K_d of 0.26 μM between DltB and DltC-Ppant (Extended Data Fig. 6). In contrast to the tight DltB–DltC interaction, DltB does not form a detectable complex with DltA or the extracellular domain of DltD, and there is no detectable interaction between DltA and DltC on the cytoplasmic side (data not shown).

To understand how DltB functions as an MBOAT, we also determined the crystal structure of the DltB–DltC-Ppant complex at 3.15 Å resolution (Fig. 2c). Cytoplasmic DltC contains four helices, with Ppant-bonded Ser35 being the first residue of helix 3 (α3). Residues of DltC α3 and the long loop between α3 and α4 (α3–α4 loop) form the DltB-binding surface. DltC interacts mainly with the C-terminal half of DltB H13 and the N-terminal end of DltB H14. This region is formed by a DltB-specific insertion that is missing in other MBOAT proteins[1]. The DltB–DltC interface is mostly hydrophobic, formed by DltB residues Met302, Val305, Ile306 and Met309, and DltC residues Met36, Val39, Val43 and Val55. In addition, Arg317—the first residue of helix H14—forms charged hydrogen bonds with DltC Glu40, whereas the phosphate group of Ser35-Ppant is in a position to form a salt bridge with DltB Lys282 in helix H12 (Fig. 2d, Extended Data Figs. 5, 7). The structures of DltB are essentially identical in both the apo state and the DltC-bound state (Extended Data Fig. 7a).

To confirm the structural and functional features of the DltB–DltC interface, we purified DltB mutants V305D and V305D/I306D as well as DltC mutants V39D and V39R, and tested their interactions with their corresponding wild-type partner using GST pull-down and Octet assays (Fig. 2b, Extended Data Fig. 6). Whereas DltB(V305D) showed substantially reduced binding to wild-type DltC, the binding was completely abolished when using DltB(V305D/I306D). Similarly, both DltC(V39D) and DltC(V39R) showed substantially reduced ability to interact with DltB. These mutagenesis analyses demonstrate that Val305 and Ile306 of DltB and Val39 of DltC are critical to the DltB–DltC interaction, and confirm our structural observation that these surface residues are located at the DltB–DltC core interface.

There is an approximately straight tunnel between the bottom of the extracellular funnel and the cytoplasmic side. This tunnel is formed by three DltB helices from the C-ridge (H13–H15) and the small horizontal helix H12 from the central core. DltB residues inside the tunnel are highly conserved among DltB proteins (Fig. 3a, Extended Data Fig. 5), and show a level of conservation in other MBOAT proteins, which suggests that this tunnel is functionally important. It should be noted that in our current structures of DltB and the DltB–DltC complex, the side chain of the conserved Trp285 from helix H12 keeps this tunnel in a closed conformation (Fig. 3a, Extended Data Fig. 7c). We speculate that the conformation we captured is that of the DltB enzyme in a 'resting' state.

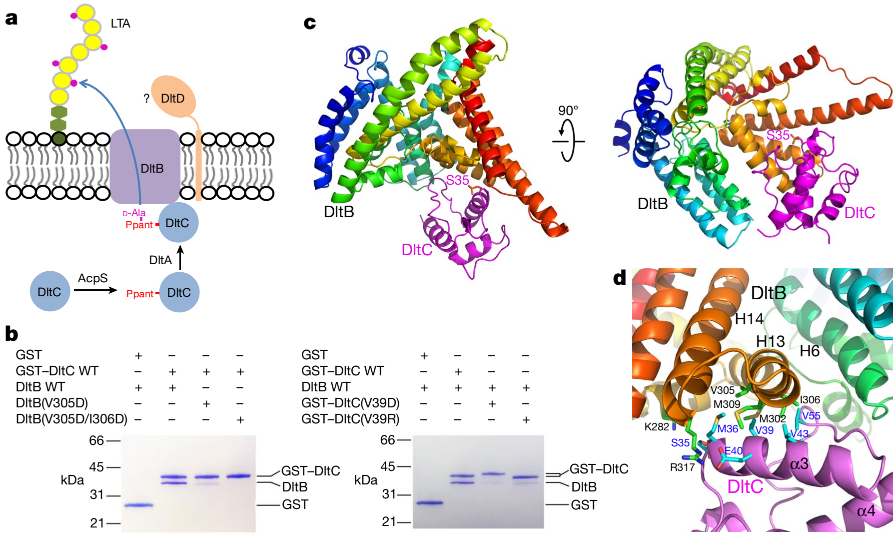


Fig. 2 | Structural basis of the DltB–DltC-Ppant interaction. a, Dlt proteins responsible for LTA D-alanylation in Gram-positive bacteria. The magenta dots on glycerophosphate units represent D-Ala moieties. **b**, Direct stable interaction between DltB and DltC-Ppant, and mutagenesis analysis of the DltB–DltC-Ppant interface, as shown by GST pull-down assays. GST pull-down experiments were performed at least twice with similar results. WT, wild type. **c**, Overall structure of the DltB–DltC-Ppant complex. DltC-Ppant binds to DltB on the cytosolic side, with the phosphate group of Ppant (which is attached through Ser35 of DltC) pointing towards the DltB tunnel. **d**, The DltB–DltC-Ppant interface. Side chains corresponding to DltB are shown as green sticks, and side chains of DltC are shown in cyan.

One notable feature of the DltB–DltC complex is that DltC Ser35 is located at the cytoplasmic entrance of the tunnel (Fig. 3b). Whereas the electron density for the Ppant phosphate group is well-defined in our electron density map, the density for the rest of the Ppant chain is too thin for model building; this is consistent with a 'resting' state conformation. Consistently, Octet analysis showed that the DltC(S35A) mutant can also interact with DltB with similar affinity ($K_d \approx 0.19\,\mu\text{M}$) to that of wild-type DltC-Ppant, which indicates that the Ppant group is not essential to the DltB–DltC interaction. The Ppant group can potentially switch between occupying the tunnel and being flexible in the cytoplasmic open space, as the Ser35 phosphate group is positioned between the tunnel entrance and the open cytoplasmic space. While the most conserved residue (His336) is located at the C terminus of the DltB H14 helix, DltC makes contacts with the C-terminal half of the H13 helix and the N terminus of the H14 helix—which suggests that the distance between DltC Ser35 and DltB His336 may be largely fixed during catalysis.

To examine the functional importance of the tunnel, we generated *B. subtilis* strains that lacked the *dlt* operon, and then complemented with a heterologous copy of the *dlt* locus expressed from its native promoter. The LTA D-alanylation level and the viability of *dlt*-deleted *B. subtilis* cells complemented with a heterologous copy of the *dlt* locus, containing either wild-type *dltB* or various *dltB* mutations, were evaluated using ^{14}C- D-Ala radiolabelling and lysozyme-sensitivity

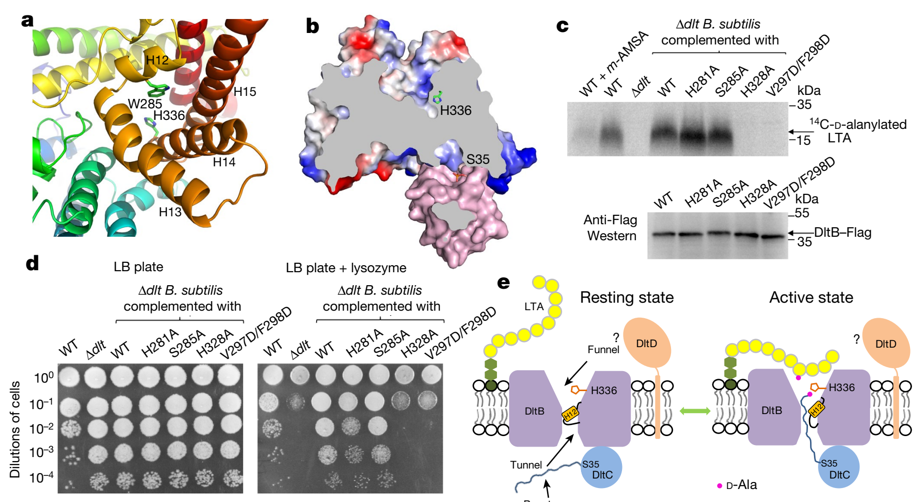


Fig. 3 | Structure of the DltB tunnel and the DltB–DltC-Ppant binding mode provide insight into the molecular mechanism of DltB.
a, Residues forming the DltB tunnel. **b**, Cut-away surface illustration of the DltB–DltC-Ppant complex. DltC-Ppant pSer35 is located at the bottom of the tunnel. **c**, LTA D-alanylation assay. *m*-AMSA is a DltB inhibitor. The assays were repeated three times. H281, S285, H328 and V297/F298 in *B. subtilis* correspond to H289, S293, H336 and V305/I306 in *S. thermophilus*, respectively. **d**, Lysozyme susceptibility survival assay. Representative images are shown for serial dilutions of cells plated on LB agar (left) and LB agar supplemented with 30 μg ml^{-1} of lysozyme (right). Dilutions of cells are indicated on the *y* axis. The survival assay was performed three times. **e**, A working model for DltB-mediated LTA D-alanylation. Cross-membrane D-alanylation is probably mediated by the DltB tunnel, the opening (activation) of which may be controlled by helix H12. The role of DltD in this reaction is unclear.

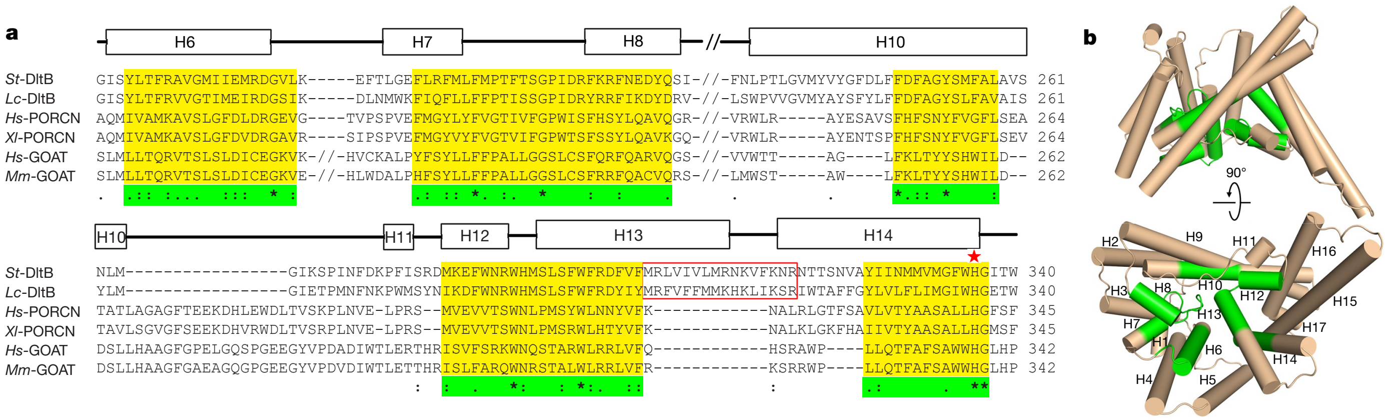


Fig. 4 | Conserved regions among bacterial DltB and vertebrate PORCN and GOAT proteins. a, Alignment of conserved regions of DltB, PORCN and GOAT. Conserved sequences are highlighted in yellow (and in green under the sequences). The red rectangle indicates the DltB-specific insertion, which is involved in the binding of DltC-Ppant. A red star marks the most conserved histidine residue among MBOATs. *St*, *S. thermophilus*; *Lc*, *Lactobacillus casei*; *Hs*, *Homo sapiens*; *Xl*, *Xenopus laevis*; *Mm*, *Mus musculus*. **b**, The conserved MBOAT core. Conserved regions among MBOATs form a central core in the DltB structure (coloured in green), whereas the peripheral helices shielding the core are largely non-conserved (coloured in wheat).

assays, respectively. Mutation of DltB residues corresponding either to *S. thermophilus* DltB His336 or to the DltC-binding site completely abolished LTA D-alanylation (Fig. 3c). In addition, both mutations considerably reduced the viability of *B. subtilis* in the presence of lysozyme, whereas mutations of two other DltB residues did not have a substantial effect in both assays (Fig. 3c, d, Extended Data Fig. 8). Our functional assay data together with the structural features of DltB strongly suggest that the tunnel is important for the catalytic mechanism of DltB (Fig. 3e).

In some other *O*-acyltransferases, such as carnitine acyltransferase[24], a conserved histidine catalyses the acyl-transfer reaction by aligning the carnitine substrate with the acyl-CoA thioester bond. The Ppant- D-Ala chain has a length of around 20 Å between the phosphate group and D-Ala. In our crystal structure, the distance between the Ser35-Ppant phosphate group and His336 is approximately 21 Å. Should the tunnel be open for the Ppant-D-Ala chain binding, this distance would enable His336 to align the substrate that receives the acyl group (probably a glycerol phosphate unit within LTA molecule) with D-alanylated DltC-Ppant. Thus, our structures suggest a model in which D-alanylation of LTA occurs between the LTA bound to the extracellular funnel and the D-alanyl group on DltC-Ppant-D-Ala bound to the cytoplasmic side of the tunnel (Fig. 3e).

Because DltB forms a stable complex with DltC-Ppant even without the D-alanyl group, and the DltC Ser35 is open to the cytosol, we speculate that DltC-Ppant forms a constitutive complex with DltB during catalysis and the Ppant chain can migrate between the tunnel and the cytosol, where loading of the D-alanyl group of DltC-Ppant can be catalysed by DltA. We then asked how the DltB tunnel opens for Ppant binding. The DltB tunnel is formed by the small horizontal helix H12 and the long transmembrane helices (H13–H15) forming the C-ridge of DltB. Compared to the DltB C-ridge helices, helix H12 is more likely to be the mobile structural component. The tunnel opening can be caused by movement or by a conformational change of the short helix H12, the position of which is stabilized through local hydrophobic interactions. H12 may change its position without disturbing the N- and C-ridge structures and lead to the opening of the tunnel. H12 movement may be induced by the presence of an appropriate signal, such as substrate binding with the extracellular funnel and/or binding of intracellular ligands such as DltC-Ppant-D-Ala. It should be noted here that in the Dlt system DltD is required for the D-alanylation of LTA in vivo. It remains unclear how DltD may contribute to this process. A combination of structural and enzymatic analysis is needed to reveal the role of DltD and the detailed catalytic mechanism of DltB.

We next considered the implication of the DltB structure for other MBOAT proteins. Despite a low overall sequence homology, a more conserved region within MBOAT sequences—termed MBOAT2 homology—was identified (http://pfam.xfam.org/family/MBOAT_2). The MBOAT2 homology covers the sequences that correspond to the DltB region from DltB H12 to the N terminus of H15 (Fig. 1b), which forms the majority of the central core that is thinner than the lipid bilayer. Thus, the thin central core and the extracellular (or lumen-facing) funnel are likely to be common structural features in many MBOATs. It has been demonstrated that the most conserved histidine (DltB His336), which is located within this MBOAT2 homology domain, is critical for the enzymatic activities of all tested MBOAT proteins—including PORCN, HHAT, GOAT, ACAT and DGAT[17–22]—which strongly suggests a common or similar catalytic mechanism for the MBOAT superfamily of proteins. The conserved extracellular/lumen structural funnel, the thin central core and the tunnel that we observed in our DltB structure are probably shared by many other MBOATs. Indeed, our crystal structure of DltB is in good agreement with the membrane topology models of HHAT and GOAT that have previously been derived from biochemical data[25–27] (Extended Data Fig. 9). For example, in each case, the catalytic histidine was predicted to be at the end of an HHAT or a GOAT transmembrane helix facing the lumen side, consistent with our structure of DltB. That the critical horizontal helices H11–H13 in our DltB core structure were predicted to be a cytoplasmic subdomain of HHAT or GOAT is also consistent with models. In addition, a predicted 're-entrant helix' observed in both HHAT and GOAT corresponds to the H7–H8 'half-way turn-back' structure in the DltB central core (Fig. 1b, Extended Data Fig. 9). In contrast to the similarity in the core structure, the N- and C-terminal regions of HHAT and GOAT are much more divergent.

Among vertebrate MBOATs, PORCN and GOAT are responsible for lipid modifications of secreted Wnts and ghrelin, respectively. They all catalyse reactions across the endoplasmic reticulum membrane, with the acyl-group-accepting proteins located in the endoplasmic reticulum lumen and acyl-CoA in the cytosol[6,7]. Because DltB also catalyses cross-membrane reactions, we examined the sequence homology among DltB, PORCN and GOAT. It appears that there are four conserved regions: the region covering DltB helices H12–H14 (the MBOAT2 homology region), the DltB H7–H8 region, and two partial helices in the inner circle of the DltB structure (most of helix H6 and the central part of helix H10) (Fig. 4, Extended Data Fig. 9). Therefore, although sequences encoding the N- and C-ridges of DltB are generally not conserved in other MBOATs, the central core of DltB—along with its structural neighbours in the inner circle (for example, parts

of helices H6 and H10)—are conserved among vertebrate MBOATs, including PORCN and GOAT. We suggest that the non-conserved nature of the ridges enable recognition of distinct substrates specific to different members within the MBOAT family. It should be noted that the mechanism of acyl-group binding is probably very different between metazoan MBOATs, most of which bind acyl-CoA as a donor, and DltB, which uses DltC-Ppant-D-Ala.

The deep, conserved DltB extracellular structural funnel, as well as the DltB tunnel, may be an excellent target for drug development. Furthermore, many other bacterial and metazoan MBOATs may also be very druggable targets, as many of them are present on the surface of the cell membrane. In addition, the deep extracellular/lumen funnel shape close to the active site is probably a conserved feature of many MBOATs, and may be an excellent drug-binding site. Indeed, even in the absence of a 3D structure and detailed enzymatic analysis using purified PORCN, multiple small-molecule inhibitors with half-maximal inhibitory concentrations in the low-nanomolar range have been found through cell-based screening, and some of them have been used in clinical trials for the treatment of cancer[28–30]. Potent HHAT and GOAT inhibitors have also been reported and examined in several studies[5]. On the basis of our crystal structures, we predict that many more highly potent MBOAT inhibitors will be discovered in the future.

Three-dimensional structural prediction of MBOAT proteins has been very difficult and unreliable. Our crystal structures of DltB serve as a cornerstone for understanding the structure and function of MBOAT proteins. In addition, our structures reveal an intriguing mechanism for cross-membrane catalysis, and provide a platform for the development of new clinically relevant drugs across species.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0568-2.

Received: 11 December 2017; Accepted: 8 August 2018;
Published online 3 October 2018.

1. Hofmann, K. A superfamily of membrane-bound *O*-acyltransferases with implications for Wnt signaling. *Trends Biochem. Sci*. **25**, 111–112 (2000).
2. Chang, T. Y., Chang, C. C., Ohgami, N. & Yamauchi, Y. Cholesterol sensing, trafficking, and esterification. *Annu. Rev. Cell Dev. Biol*. **22**, 129–157 (2006).
3. Liu, Q., Siloto, R. M., Lehner, R., Stone, S. J. & Weselake, R. J. Acyl-CoA:diacylglycerol acyltransferase: molecular biology, biochemistry and biotechnology. *Prog. Lipid Res*. **51**, 350–377 (2012).
4. Chang, S. C. & Magee, A. I. Acyltransferases for secreted signalling proteins (review). *Mol. Membr. Biol*. **26**, 104–113 (2009).
5. Masumoto, N. et al. Membrane bound *O*-acyltransferases and their inhibitors. *Biochem. Soc. Trans*. **43**, 246–252 (2015).
6. Resh, M. D. Fatty acylation of proteins: the long and the short of it. *Prog. Lipid Res*. **63**, 120–131 (2016).
7. Tuladhar, R. & Lum, L. Fatty acyl donor selectivity in membrane bound *O*-acyltransferases and communal cell fate decision-making. *Biochem. Soc. Trans*. **43**, 235–239 (2015).
8. Resh, M. D. Palmitoylation of proteins in cancer. *Biochem. Soc. Trans*. **45**, 409–416 (2017).
9. Lanyon-Hogg, T., Faronato, M., Serwa, R. A. & Tate, E. W. Dynamic protein acylation: new substrates, mechanisms, and drug targets. *Trends Biochem. Sci*. **42**, 566–581 (2017).
10. Madan, B. & Virshup, D. M. Targeting Wnts at the source—new mechanisms, new biomarkers, new drugs. *Mol. Cancer Ther*. **14**, 1087–1094 (2015).
11. Perego, M. et al. Incorporation of D-alanine into lipoteichoic acid and wall teichoic acid in *Bacillus subtilis*. Identification of genes and regulation. *J. Biol. Chem*. **270**, 15598–15606 (1995).
12. Neuhaus, F. C., Heaton, M. P., Debabov, D. V. & Zhang, Q. The *dlt* operon in the biosynthesis of D-alanyl-lipoteichoic acid in *Lactobacillus casei*. *Microb. Drug Resist*. **2**, 77–84 (1996).
13. Reichmann, N. T., Cassona, C. P. & Gründling, A. Revised mechanism of D-alanine incorporation into cell wall polymers in Gram-positive bacteria. *Microbiology* **159**, 1868–1877 (2013).
14. Pasquina, L. et al. A synthetic lethal approach for compound and target identification in *Staphylococcus aureus*. *Nat. Chem. Biol*. **12**, 40–45 (2016).
15. Neuhaus, F. C. & Baddiley, J. A continuum of anionic charge: structures and functions of D-alanyl-teichoic acids in Gram-positive bacteria. *Microbiol. Mol. Biol. Rev*. **67**, 686–723 (2003).
16. Peschel, A. et al. Inactivation of the *dlt* operon in *Staphylococcus aureus* confers sensitivity to defensins, protegrins, and other antimicrobial peptides. *J. Biol. Chem*. **274**, 8405–8410 (1999).
17. Rios-Esteves, J., Haugen, B. & Resh, M. D. Identification of key residues and regions important for porcupine-mediated Wnt acylation. *J. Biol. Chem*. **289**, 17009–17019 (2014).
18. Yang, J., Brown, M. S., Liang, G., Grishin, N. V. & Goldstein, J. L. Identification of the acyltransferase that octanoylates ghrelin, an appetite-stimulating peptide hormone. *Cell* **132**, 387–396 (2008).
19. Buglino, J. A. & Resh, M. D. Identification of conserved regions and residues within hedgehog acyltransferase critical for palmitoylation of sonic hedgehog. *PLoS ONE* **5**, e11195 (2010).
20. Das, A., Davis, M. A. & Rudel, L. L. Identification of putative active site residues of ACAT enzymes. *J. Lipid Res*. **49**, 1770–1781 (2008).
21. Lin, S., Lu, X., Chang, C. C. & Chang, T. Y. Human acyl-coenzyme A:cholesterol acyltransferase expressed in chinese hamster ovary cells: membrane topology and active site location. *Mol. Biol. Cell* **14**, 2447–2460 (2003).
22. McFie, P. J., Stone, S. L., Banman, S. L. & Stone, S. J. Topological orientation of acyl-CoA:diacylglycerol acyltransferase-1 (DGAT1) and identification of a putative active site histidine and the role of the N terminus in dimer/tetramer formation. *J. Biol. Chem*. **285**, 37377–37387 (2010).
23. Viana, D. et al. A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat. Genet*. **47**, 361–366 (2015).
24. Jogl, G., Hsiao, Y. S. & Tong, L. Structure and function of carnitine acyltransferases. *Ann. NY Acad. Sci*. **1033**, 17–29 (2004).
25. Taylor, M. S. et al. Architectural organization of the metabolic regulatory enzyme ghrelin *O*-acyltransferase. *J. Biol. Chem*. **288**, 32211–32228 (2013).
26. Matevossian, A. & Resh, M. D. Membrane topology of hedgehog acyltransferase. *J. Biol. Chem*. **290**, 2235–2243 (2015).
27. Konitsiotis, A. D. et al. Topological analysis of hedgehog acyltransferase, a multipalmitoylated transmembrane protein. *J. Biol. Chem*. **290**, 3293–3307 (2015).
28. Barnett, B. P. et al. Glucose and weight control in mice with a designed ghrelin *O*-acyltransferase inhibitor. *Science* **330**, 1689–1692 (2010).
29. Ho, S. Y. & Keller, T. H. The use of porcupine inhibitors to target Wnt-driven cancers. *Bioorg. Med. Chem. Lett*. **25**, 5472–5476 (2015).
30. Chen, B. et al. Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. *Nat. Chem. Biol*. **5**, 100–107 (2009).

Acknowledgements We are grateful to the staff at Advanced Light Source beamlines 5.0.1, 8.2.1 and 8.2.2 for assistance with synchrotron data collection. We thank N. Zheng and P. Hsu for comments on this manuscript, T. Hinds for discussion and advice on assays, S. Ovchinnikov for computational modelling, L. Kruse for use of the radioactive gel scanner and M. Ragheb for construction of the *dlt* operon deletion strain. This work was supported by National Institutes of Health grant R01 GM127316 to W.X. and a Jane Coffin Childs postdoctoral fellowship to D.M. This work was also supported by Chinese Academy of Sciences grant XDB08010303 to Z.R. and W.X., the National Institute of Health grant DP2GM110773 to H.M. and the Bacterial Pathogenesis Training Grant 5T32AI055396-13 to K.S.L.

Reviewer information *Nature* thanks E. Tate and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.M. carried out protein purification, crystallization, and related binding and enzymatic analysis. D.M. and Z.W. collected diffraction data and determined the crystal structure. Z.W. performed structural refinement. C.N.M. constructed the *B. subtilis* strains, and C.N.M., K.S.L. and H.M. performed the cell survival assays. P.L. contributed to molecular cloning and sample preparation. X.L. and Z.R. contributed to screening of other MBOAT proteins. D.M., Z.W. and W.X. analysed structural data and wrote the paper. All authors participated in manuscript revision and analysis of biochemical data.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0568-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0568-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to W.X.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Protein preparation. The cDNA of full-length *S. thermophilus* DltB was subcloned into pET21b (Novagen). cDNAs of *S. thermophilus* AcpS, DltA and DltC were subcloned into pQLink vectors (Addgene) with AcpS and DltA bearing an N-terminal 6×His-tag, and DltC bearing a N-terminal GST-tag. *Escherichia coli* strain C43 (DE3) was used for protein overexpression. Overexpression of the above proteins was induced by 0.4 mM isopropyl β-D-thiogalactoside when cell density reached an optical density at 600 nm (OD_{600}) of 1.0. After induction at 37 °C for 5 h, the cells were collected and homogenized in buffer containing 25 mM Tris-HCl pH 8.0 and 150 mM NaCl.

For DltB purification, after disruption by sonication, cell debris was removed by centrifugation for 10 min at 20,000*g*. The supernatant was collected and ultracentrifuged for 1.5 h at 100,000*g*. The membrane fraction was collected and homogenized with buffer containing 25 mM Tris-HCl pH 8.0 and 150 mM NaCl. *n*-Decyl-β-D-maltopyranoside (Anatrace) was added to the membrane suspension to a final concentration of 1.5% (w/v) and then incubated for 2 h at 4 °C. After another ultracentrifugation step at 100,000*g* for 30 min, the supernatant was collected and loaded onto Ni-NTA affinity resin (Ni-NTA; Qiagen). After washing with buffer containing 25 mM Tris-HCl pH 8.0, 500 mM NaCl, 25 mM imidazole and 0.2% (w/v) *n*-decyl-β-D-maltopyranoside, DltB was eluted with a buffer containing 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 400 mM imidazole and various detergents from Anatrace. After being concentrated to 10 mg ml^{-1}, DltB was further purified by gel filtration (Superdex-200 10/30; GE Healthcare). The buffer for gel filtration contained 25 mM Tris-HCl pH 8.0, 150 mM NaCl and various detergents. The peak fractions were collected.

For the purification of DltA and DltC, after sonication the cell debris was removed by centrifugation for 1 h at 35,000*g*. The supernatants were loaded onto Ni-NTA affinity resin and Glutathione Sepharose 4 resin (GS4B resin, GE Healthcare), respectively. After a wash step, the N-terminal GST-tag was either removed from DltC or maintained, depending on the purpose of the experiment. After elution, DltA and DltC solutions were loaded onto HiTrap Q HP columns (5 ml, GE Healthcare), and protein samples eluted from the Q column were further purified by gel filtration. Peak fractions were collected and concentrated. Finally, DltA and DltC were stored in buffer containing 25 mM Tris-HCl pH 8.0 and 150 mM NaCl.

DltB and DltC mutants were generated with a standard PCR-based strategy and were subcloned, overexpressed and purified in the same way as the wild-type proteins.

Protein crystallization. The hanging-drop vapour-diffusion method was performed at room temperature during crystallization. DltB and DltC proteins were purified as mentioned above, and crystals were obtained from DltB purified with *n*-nonyl-β-D-glucopyranoside (Anatrace). For crystallization of the DltB–DltC complex, DltB and DltC were purified separately and mixed before crystallization at a molar ratio of 1:2. Crystals belonging to crystal form I (space group $P2_1$, Extended Data Table 1) were crystallized in buffer containing 21% PEG400, 100 mM Tris-HCl pH 7.5, 100 mM NaCl and 100 mM $MgCl_2$. Crystals belonging to crystal form II (space group $P2_1$, Extended Data Table 1) were crystallized in buffer containing 27% PEG400, 100 mM sodium citrate pH 5.6, 200 mM $NH_4H_2PO_4$ and 100 mM $(NH_4)_2SO_4$. Crystals belonging to crystal form III (space group $P2_12_12_1$, Extended Data Table 1) were crystallized in buffer containing 27% PEG400, 100 mM HEPES pH 7.5, 200 mM sodium citrate tribasic dihydrate and 3% 1,5-diaminopentane dihydrochloride. For crystals in the different crystal forms above, thin or thick rod-shaped crystals typically grew for 1 to 2 weeks before reaching full crystal size. Gold derivatives were obtained by soaking the crystals in crystal form I for 2 h in mother liquor containing 2 mg ml^{-1} $KAu(CN)_2$.

Data collection and structure determination. The crystals were directly flash-frozen in liquid nitrogen. Screening and data collection were performed at the Advanced Light Source, beamlines 5.0.1, 8.2.1 and 8.2.2. All diffraction data were processed by HKL2000[31]. The single-wavelength anomalous dispersion (SAD) dataset was collected near the gold L-III absorption edge at a wavelength of 1.02 Å (Extended Data Table 1). The gold derivative sites and the initial phases were determined by PHENIX[32]. Twenty gold derivative sites were found in one asymmetric unit, and the experimental electron density map clearly showed the presence of four DltB molecules in one asymmetric unit. The *B. subtilis* DltC crystal structure (PDB ID: 4BPH) was used as the searching model for our DltC molecules[33]. The complex model was improved using iterative cycles of manual rebuilding with the program Coot[34] and refinement with a native dataset of 3.30 Å using Refmac5 of the CCP4 7.0 program suite[35]. The structures for crystal forms II and III were solved by molecular replacement using the model from crystal form I. All structure model figures in the paper were generated using PyMOL[36]. The protein conservation surface was generated using the ConSurf server[37], based on the alignment of DltB sequences generated using T-Coffee[38].

Binding assay. Pull-down assays were performed as described below. Twenty micrograms of wild-type DltB (or DltB mutants), 10 μg of wild-type GST–DltC (or GST–DltC mutants) and 10 μl GS4B resin were mixed in 100 μl of pull-down buffer containing 25 mM HEPES pH 7.5, 150 mM NaCl and 0.15% (w/v) *n*-decyl-β-D-maltopyranoside. The mixed samples were incubated at 4 °C on a rotisserie for 1 h, followed by washing the resin with pull-down buffer three times. During each wash, 100 μl of pull-down buffer was added to each sample and the solution was incubated at room temperature for 2 min before centrifugation and removal of supernatant. After washing, the resin samples were analysed by SDS–PAGE with Coomassie blue staining.

Binding assays were also performed at room temperature using the Octet system (FortéBio). Free GST, and GST-tagged wild-type DltC or DltC mutants were mobilized on anti-GST biosensors (FortéBio). After quenching with free GST to block free antibody sites on the biosensors, the biosensors were dipped into DltB solutions for binding measurements. The concentration gradient of DltB used in the Octet binding assay is: 0.03 μM, 0.1 μM, 0.3 μM, 1 μM, 3 μM, 10 μM.

Construction of *B. subtilis* strain for functional assays. The *cat* gene was amplified by PCR from pGEMcat, and 500 bp upstream and downstream of *dlt* operon fragments were amplified from the *B. subtilis* genome. These three pieces were assembled using isothermal assembly and transformed directly into the *B. subtilis* HM1 strain, resulting in *dlt*-operon-deleted *B. subtilis* (Δ*dlt*). The deletion was confirmed by PCR amplification and Sanger sequencing.

The natural *dlt* locus was amplified and cloned into pMMB752. Mutations of the *dltB* gene in pMMB752 carrying the *dlt* operon and Flag-tagged constructs were generated on the basis of a standard PCR method, followed by isothermal assembly to ligate the ends together. The pMMB752 constructs were transformed into *B. subtilis* with the *dlt* operon deleted from its native locus to generate strains for use in assays. Cells used here and in the following functional experiments were cultured in the presence of appropriate antibiotics to avoid possible contamination.

Detection of LTA D-alanylation. This assay was established on the basis of a previously reported method[14]. Wild-type *B. subtilis* HM1 strain, and *dlt*-operon-deleted *B. subtilis* HM1 strain complemented with either empty pMMB752 vector or vectors containing natural *dlt*-operon-bearing mutations on the *dltB* gene (untagged or Flag-tagged), were inoculated from fresh colonies on plate into liquid LB medium supplemented with 0.5 μg ml^{-1} erythromycin. Overnight cultures were diluted into 3 ml of LB at an OD_{600} of 0.1 and grown to an OD_{600} of 0.6. Cells were pelleted and resuspended into 1.5 ml of assay medium containing 0.25× LB, 50 mM Bis-Tris pH 6.0, and 200 μg ml^{-1} D-cycloserine. To test the inhibition of *m*-AMSA on LTA D-alanylation for wild-type *B. subtilis*, a final concentration of 150 μM of *m*-AMSA (Abcam) was supplemented into the assay medium. After incubation in the assay medium for 30 min, ^{14}C-D-alanine (Moravek Biochemicals) was added to a final concentration of 25 μM for an additional incubation of 30 min or 120 min. Cells were pelleted and resuspended with SDS-loading buffer, followed by a freeze–thaw cycle. Samples were vortexed and boiled for 5 min before loading onto 4–20% gradient Tris/glycine gel (Bio-Rad). Gels were dried and exposed to a phosphor storage screen for 3 days before imaging with Typhoon FLA 9000 gel imaging scanner (GE Healthcare).

To compare the expression level of C-terminal Flag-tagged DltB in corresponding *B. subtilis* strains, each strain was cultured in 1 l LB to OD_{600} of 0.6. Cells were collected and disrupted by French press, and the cell membrane was isolated by ultracentrifugation after removing cell debris by low-speed centrifugation. The membrane of each strain was resuspended with buffer containing 25 mM Tris-HCl 8.0, 150 mM NaCl into 500 μl, followed by freezing at −80 °C. One microlitre of each membrane sample was run onto SDS–PAGE and the expression of Flag-tagged DltB was detected by western blotting.

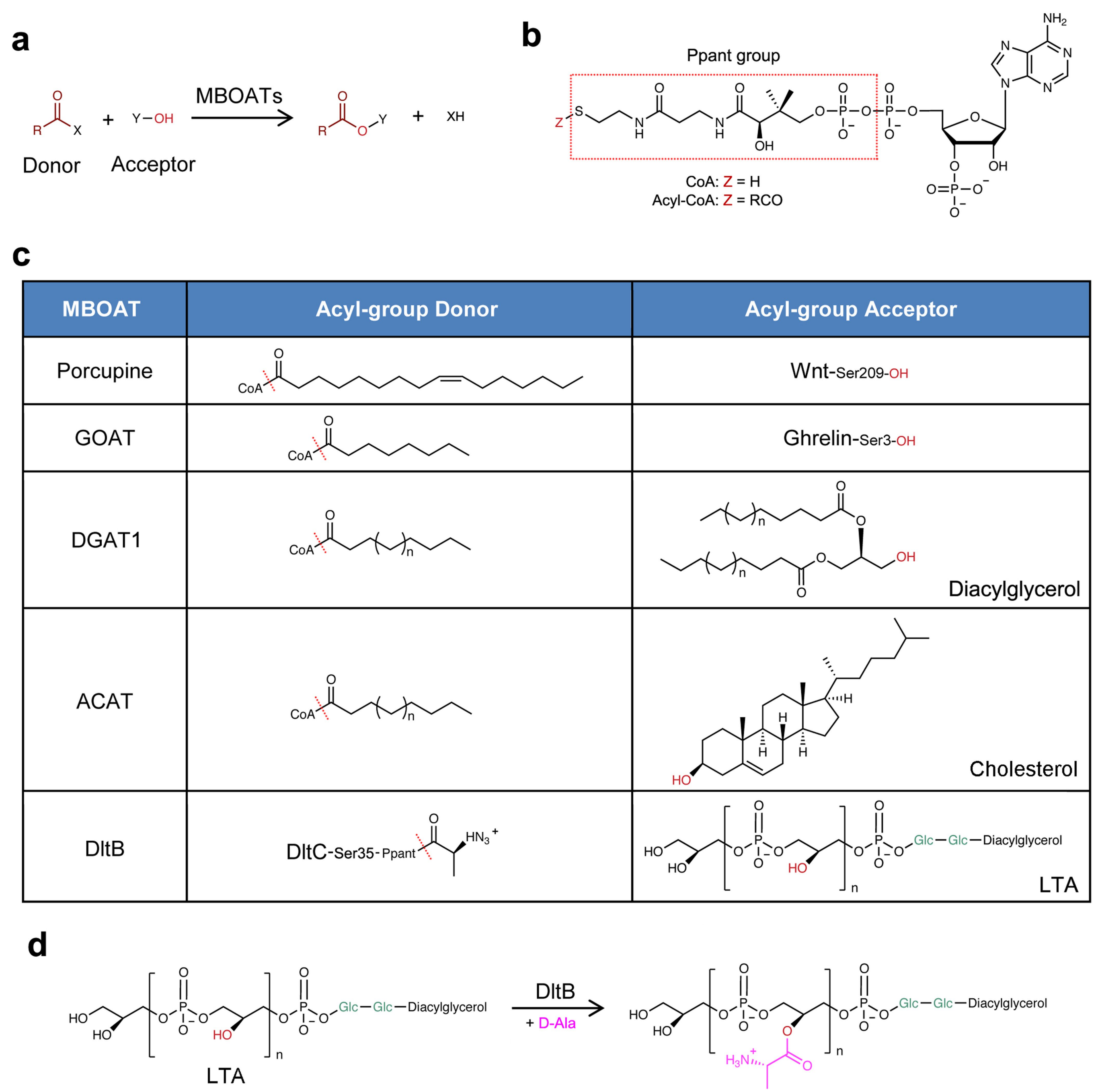
Survival assays. *Dlt* knockout strains of the Gram-positive bacterium *B. subtilis* are sensitive to the cell-wall-degrading enzyme lysozyme[39]. *B. subtilis* strains were struck on LB plates (supplemented with the appropriate antibiotic when needed) from freezer stocks and incubated overnight at 37 °C. The resulting growth on plates was used to inoculate 2-ml LB broth cultures in glass tubes. The cultures were grown at 37 °C with shaking (260 r.p.m.) to an OD_{600} of 1.0–2.0. All of the cultures were adjusted to an OD_{600} of 0.3 and then serially diluted in LB broth with tenfold dilutions. For each strain, 5 μl of each dilution was plated onto LB plates and LB plates supplemented with 30 μg ml^{-1} of lysozyme (Fisher) and incubated at 30 °C overnight. After incubation, colonies were enumerated and plates were imaged with a Bio-Rad Gel Doc XR+ Molecular Imager.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

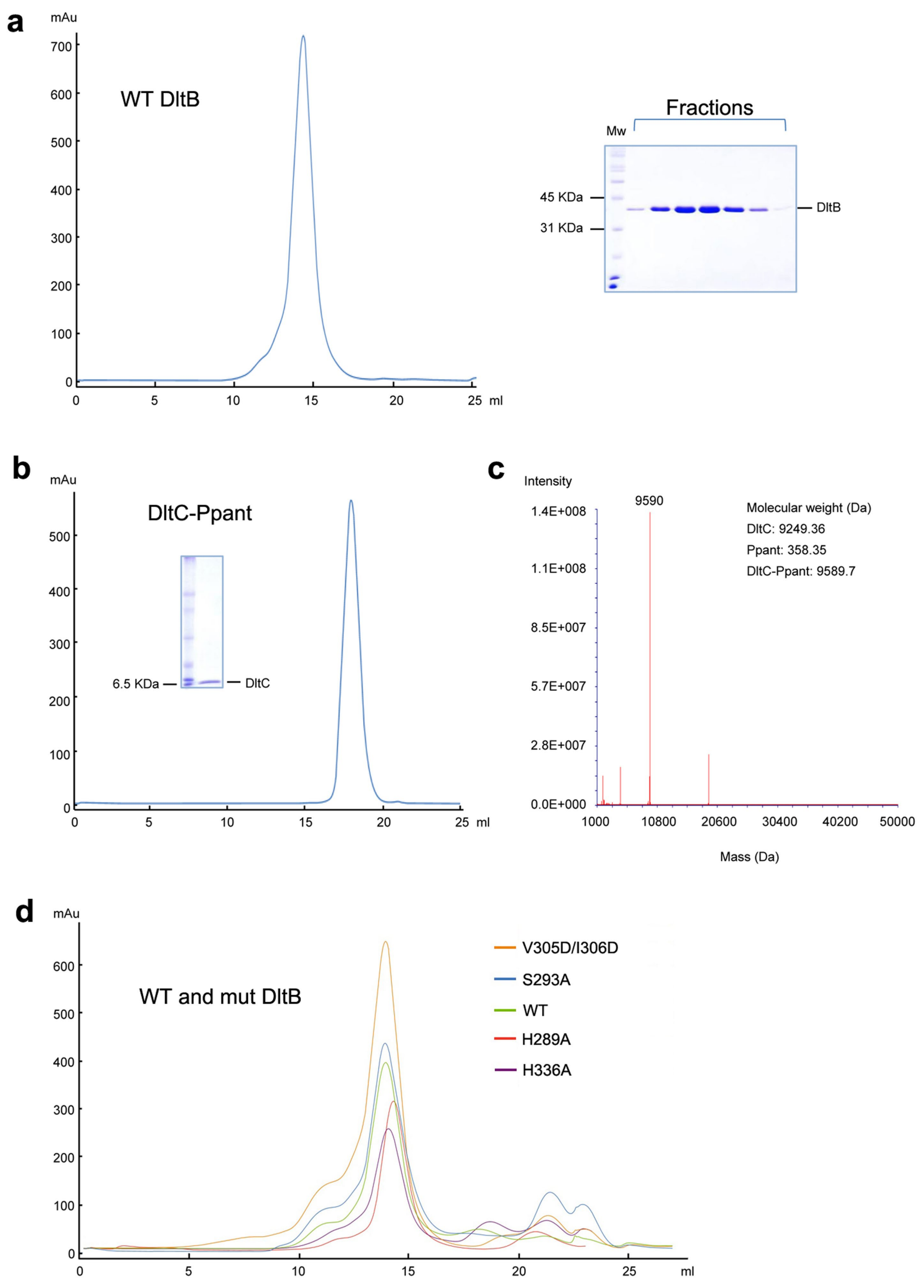
Data availability

Atomic structures have been deposited in the Protein Data Bank (PDB) with accession codes 6BUG (crystal form I), 6BUH (crystal form II) and 6BUI (crystal form III). All other data that support the findings of this study are available from the corresponding author upon reasonable request.

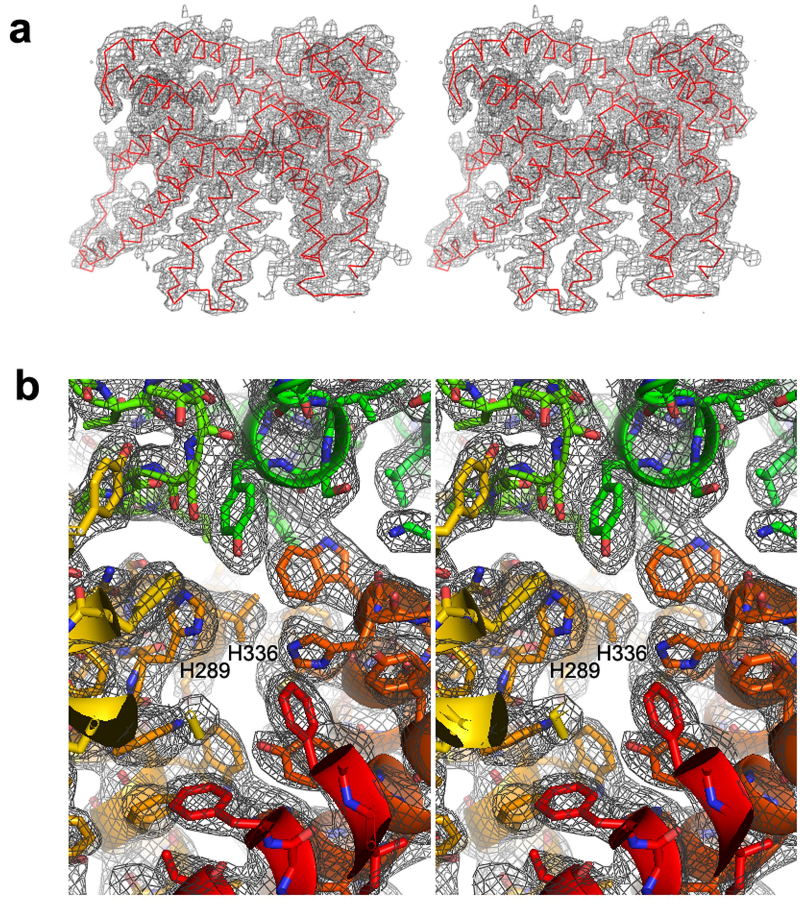
31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol*. **276**, 307–326 (1997).
32. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
33. Zimmermann, S. et al. High-resolution structures of the D-alanyl carrier protein (Dcp) DltC from *Bacillus subtilis* reveal equivalent conformations of apo- and holo-forms. *FEBS Lett*. **589**, 2283–2289 (2015).
34. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
35. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
36. Delano, W. L. & Brünger, A. T. Helix packing in proteins: prediction and energetic analysis of dimeric, trimeric, and tetrameric GCN4 coiled coil structures. *Proteins* **20**, 105–123 (1994).
37. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*. **44**, W344–W350 (2016).
38. Taly, J. F. et al. Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. *Nat. Protoc*. **6**, 1669–1682 (2011).
39. Guariglia-Oropeza, V. & Helmann, J. D. *Bacillus subtilis* σ^V confers lysozyme resistance by activation of two cell wall modification pathways, peptidoglycan *O*-acetylation and D-alanylation of teichoic acids. *J. Bacteriol*. **193**, 6223–6232 (2011).



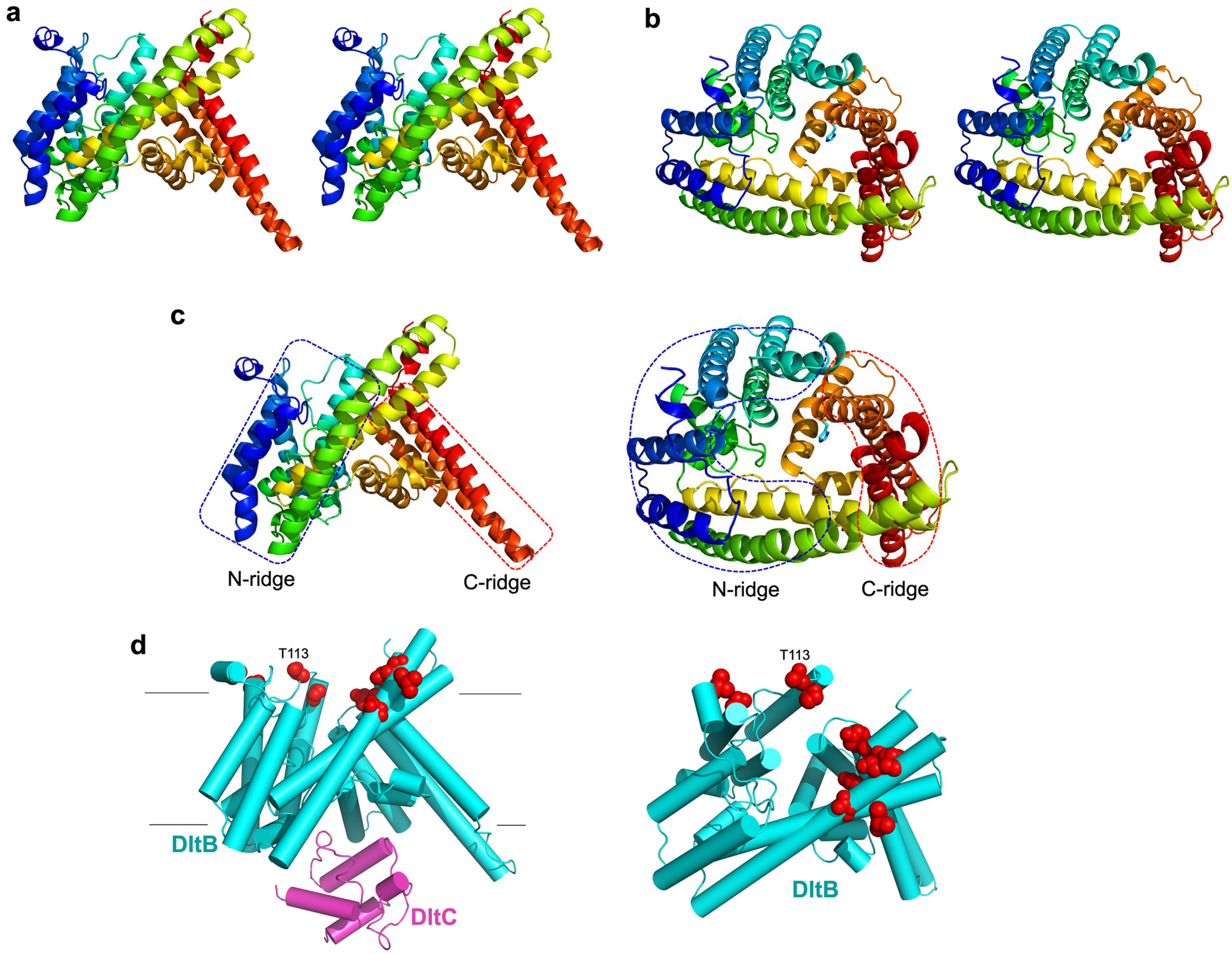
Extended Data Fig. 1 | MBOAT-catalysed reactions and chemical structures of MBOAT substrates. a, General reaction catalysed by MBOATs. **b**, Structure of CoA and acyl-CoA. The red rectangle highlights the Ppant prosthetic group within the CoA structure. For known acyl-group donors of MBOATs, the acyl groups are covalently linked with a sulfhydryl group (for example, that of Ppant in acyl-CoA or DltC-Ppant). **c**, Comparison of acyl-group donors and acceptors of PORCN, GOAT, DGAT1, ACAT and DltB. In the acyl-group donor column, the red dashed lines indicate the bonds that are broken during acyl-transfer reactions. In the acyl-group acceptor column, the hydroxyl groups that accept acyl groups are highlighted in red. ACAT1, ACAT2 and DGAT1 use saturated and unsaturated long-chain acyl-CoA. **d**, The reaction catalysed by DltB. DltB catalyses D-alanylation of both wall teichoic acid and LTA. Because the D-alanylation of wall teichoic acid is at least partially dependent on LTA D-alanylation, here we discuss only the D-alanylation of LTA. DltB transfers D-alanyl groups onto hydroxyl groups of the polyglycerolphosphate chain of the LTA molecule. For simplicity, only the type I LTA structure is shown here. The fatty-acid chains are responsible for the anchoring of LTA to the membrane of Gram-positive bacteria.



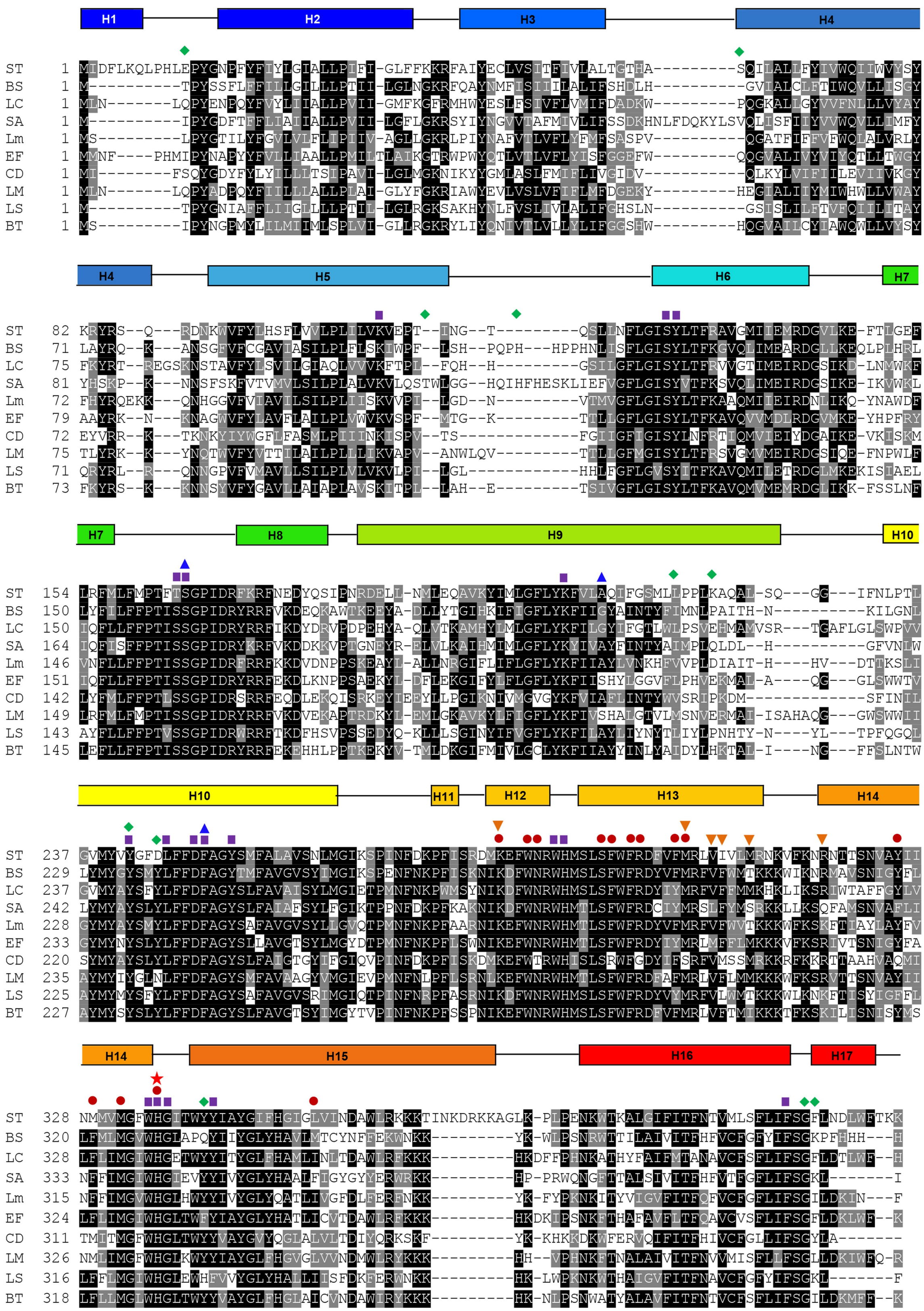
Extended Data Fig. 2 | Purification of DltB, DltC-Ppant and DltB mutants. a, SEC profile of DltB. DltB can be purified to homogeneity in most detergents and is well-behaved during SEC. **b**, SDS–PAGE and SEC profile of DltC. **c**, Mass spectrometry analysis of DltC species. This indicates that purified DltC has a molecular mass of 9,590 Da, which is equal to the calculated molecular mass of Ppant-modified DltC, referred to as DltC-Ppant. **d**, SEC profile of wild-type and mutant DltB proteins. DltB mutants including V305D/I306D, S293A, H289A and H336A are properly folded, as they migrate predominantly as a monomeric peak, similar to wild-type DltB.



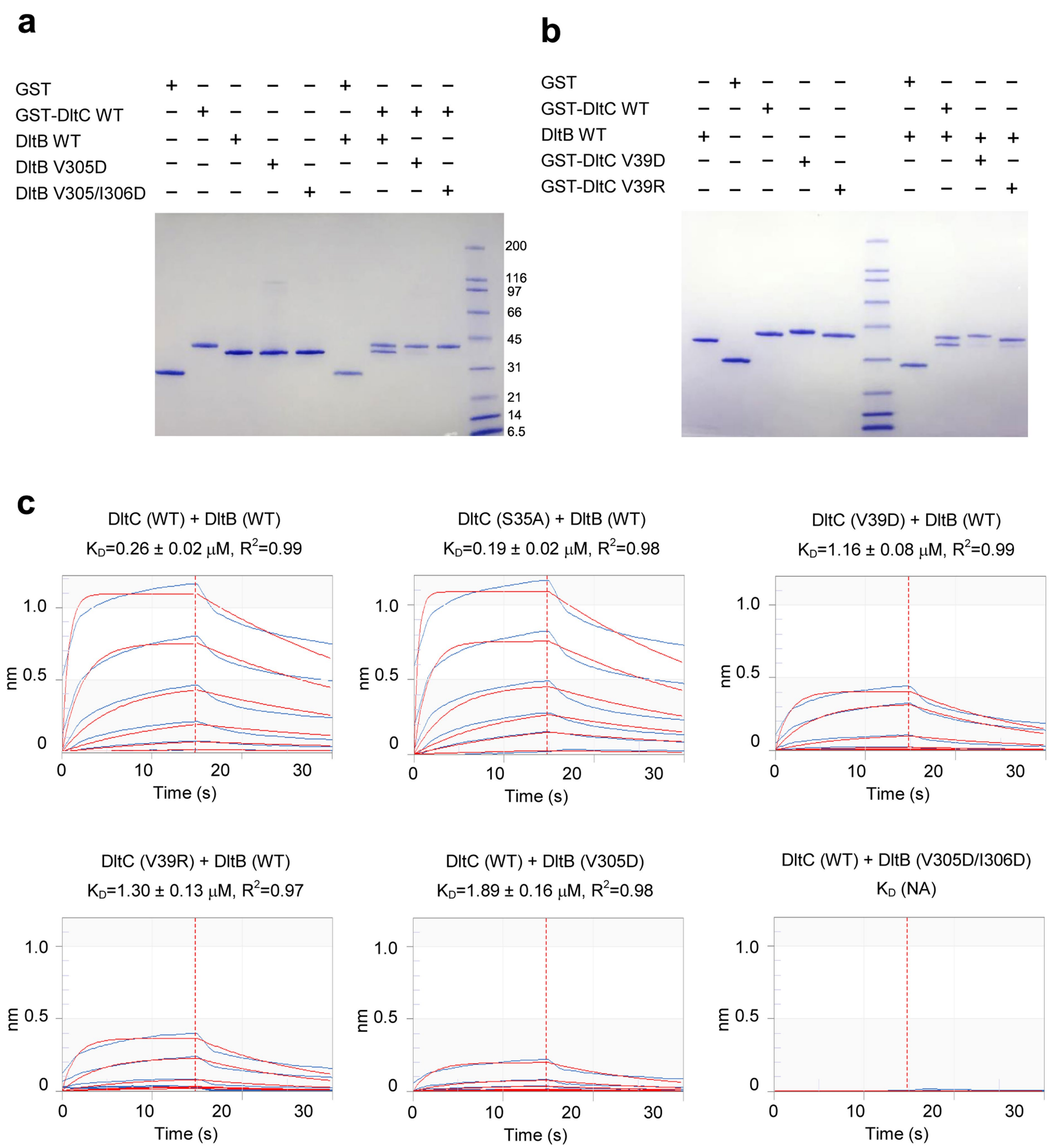
Extended Data Fig. 3 | Electron density map of DltB. a, Stereo experimental electron density map, using phases derived from an Au-SAD phasing (Extended Data Table 1). This $2F_o–F_c$ map is contoured at 1.0σ. DltB backbone tracing is shown in red. **b**, The final $2F_o–F_c$ electron density map of the crystal form II (Extended Data Table 1). This map is contoured at 1.0σ, shown in stereo and in an orientation approximately looking down the funnel. The catalytic His336 as well as His289 (another conserved residue (either His or Asn) among MBOAT proteins) are labelled. Both His336 and His289 are located at the bottom of the extracellular funnel, and sandwich the top opening of the transmembrane tunnel.



Extended Data Fig. 4 | Stereo view of DltB structure, and an extracellular 'ring' of DltB residues associated with a switch of pathogen host. a, The 'front' side view of DltB (stereo view is provided). **b**, The 'top' view of DltB, looking from the extracellular space (stereo view is provided). The His336 side chain is shown as sticks. The extracellular funnel is clear at this angle. **c**, Cartoon illustration of the N- and C-ridges of DltB in two orthogonal views. **d**, Locations of pathogen-host-sensitive sites in *S. aureus* DltB (I2, V61, T113, H121, I227, Q231, Y247, Y250, Y346, G401 and K402) are labelled with red balls in corresponding residues of the *S. thermophilus* DltB structure. It is clear that all 11 sites are located at the apex of the extracellular ridge of DltB. *S. aureus* DltB T113 is not conserved and does not have a corresponding residue in other DltBs (see Extended Data Fig. 5): here, the position of its closest residue is labelled. The intracellular DltC is shown in magenta. The DltB structure in these two panels are related with a 45° rotation.



Extended Data Fig. 5 | DltB sequence alignment. DltB sequences of representatives from 10 different genera of Gram-positive bacteria were chosen for sequence alignment using the T-Coffee server. Secondary structural elements of DltB are indicated above the alignment. Residues that form the funnel are identified by purple squares, and residues that form the tunnel are identified with dark red dots. DltB residues involved in direct interaction with DltC are indicated with orange inverted triangles. Residues corresponding to the three sites for which single-point mutations desensitize *S. aureus* to inhibition by *m*-AMSA are indicated with blue triangles. Residues of *S. aureus* DltB, the mutation of which alter the host preference from being human-specific to being capable of infecting rabbits, are indicated with green diamonds. A red star highlights the histidine residue that is completely conserved among MBOATs. ST, *S. thermophilus*; BS, *B. subtilis*; LC, *L. casei*; SA, *S. aureus*; Lm, *Listeria monocytogenes*; EF, *Enterococcus faecalis*; CD, *Clostridioides difficile*; LM, *Leuconostoc mesenteroides*; LS, *Lysinibacillus sphaericus*; BT, *Brochothrix thermosphacta*.

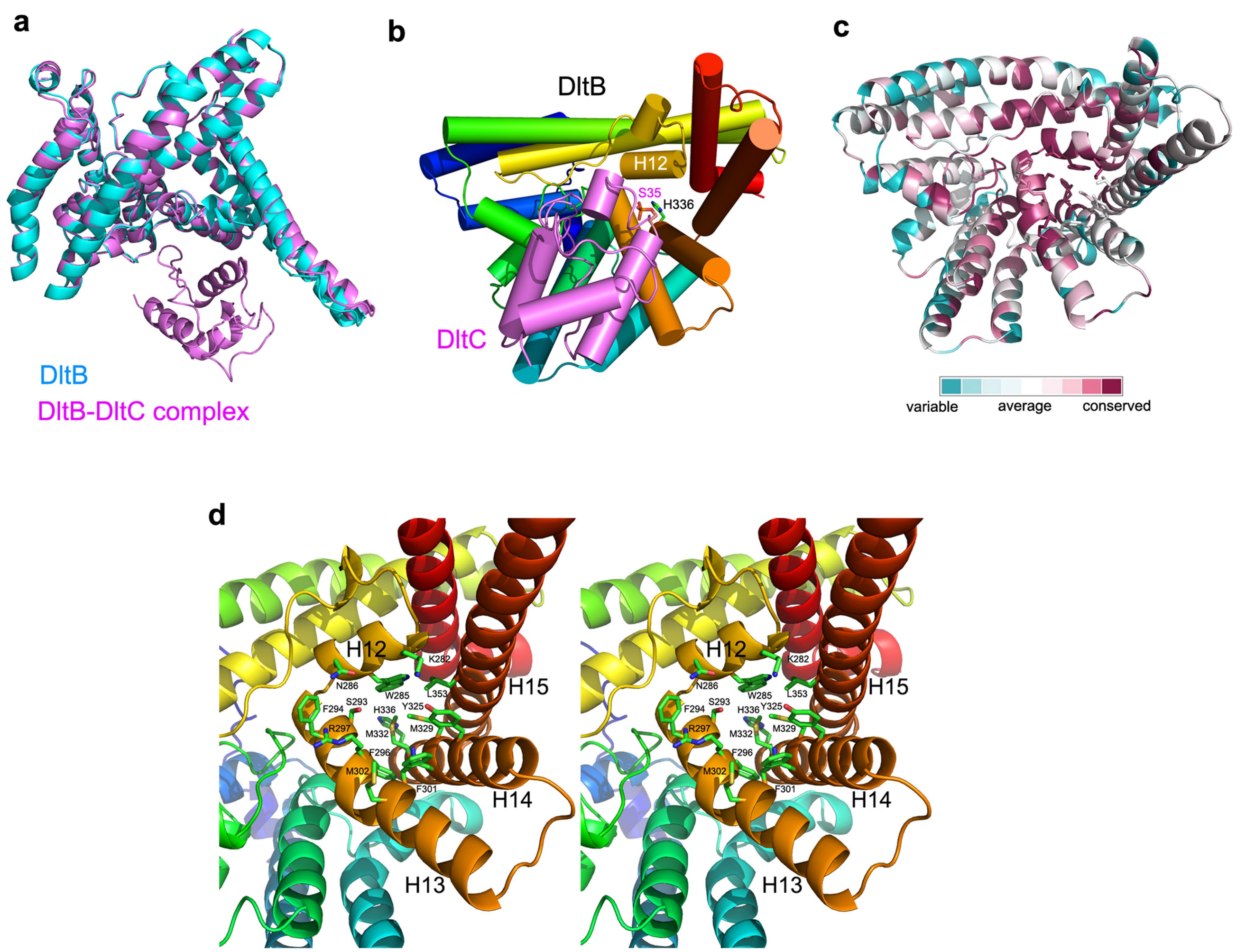


d

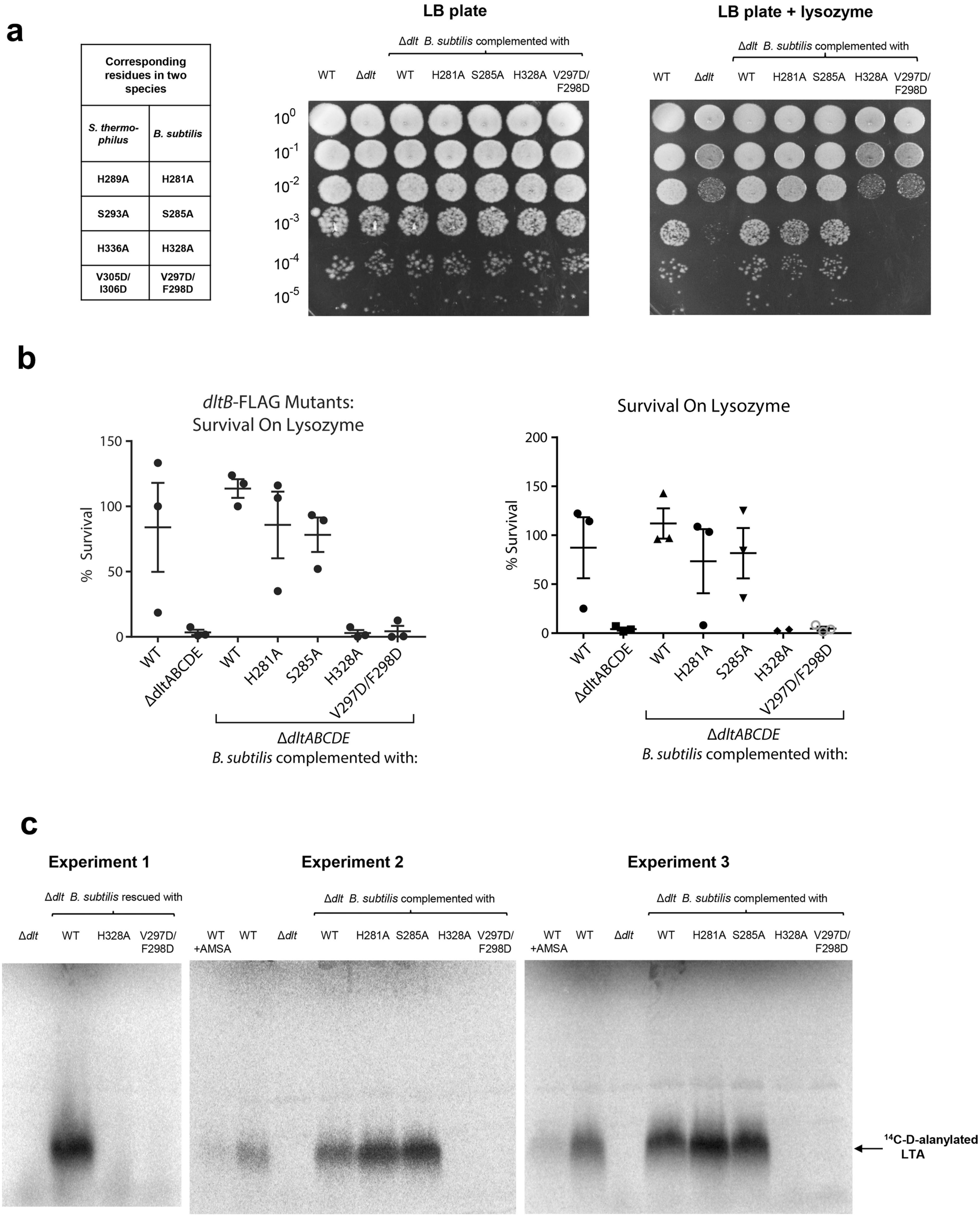
Samples in Octet Assay		K_D (μM)
GST-DltC	DltB	
WT	WT	0.26 ± 0.02
S35A	WT	0.19 ± 0.02
V39D	WT	1.16 ± 0.08
V39R	WT	1.30 ± 0.13
WT	V305D	1.89 ± 0.16
WT	V305D/I306D	NA

Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | GST pull-down and Octet assays for analysis of the interaction between DltB and DltC-Ppant. a, Results of using wild-type GST–DltC to pull-down either wild-type or mutant DltB, with GST to pull-down wild-type DltB as a negative control. Lanes 1–5 show inputs in this experiment. Pull-down results demonstrate that DltB and DltC can form a stable complex at an almost 1:1 molar ratio. DltB(V305D) loses most of its capacity to bind to wild-type GST–DltC, whereas the binding between DltB and DltC was completely abolished with the double mutant DltB(V305D/I306D). **b**, Results of using wild-type or mutant GST–DltC to pull-down wild-type DltB. Lanes 1–5 show inputs in this experiment. The mutant GST–DltC(V39D) runs slightly slower than wild-type GST–DltC and GST–DltC(V39R) on SDS–PAGE. Both GST–DltC(V39D) and GST–DltC(V39R) lost most of their capacity to bind with wild-type DltB. Pull-down experiments were performed at least twice technically, with the same results. **c**. Binding-affinity measurements for DltB and DltC using the Octet technique. Wild-type GST–DltC-Ppant and GST–DltC(S35A) show similar binding affinities with wild-type DltB. Data are shown in blue, with the corresponding fits in red. The DltB concentration gradient used here is: 0.03 μM, 0.1 μM, 0.3 μM, 1 μM, 3 μM, 10 μM. Octet assays were performed twice technically. **d**, Summary of Octet binding assay. Wild-type DltC and GST–DltC(S35A) show similar binding affinities to wild-type DltB. Mean K_d values and s.d. are shown for each assay. Mutation of residues on the binding surface of either DltB or DltC can reduce or abolish their binding.

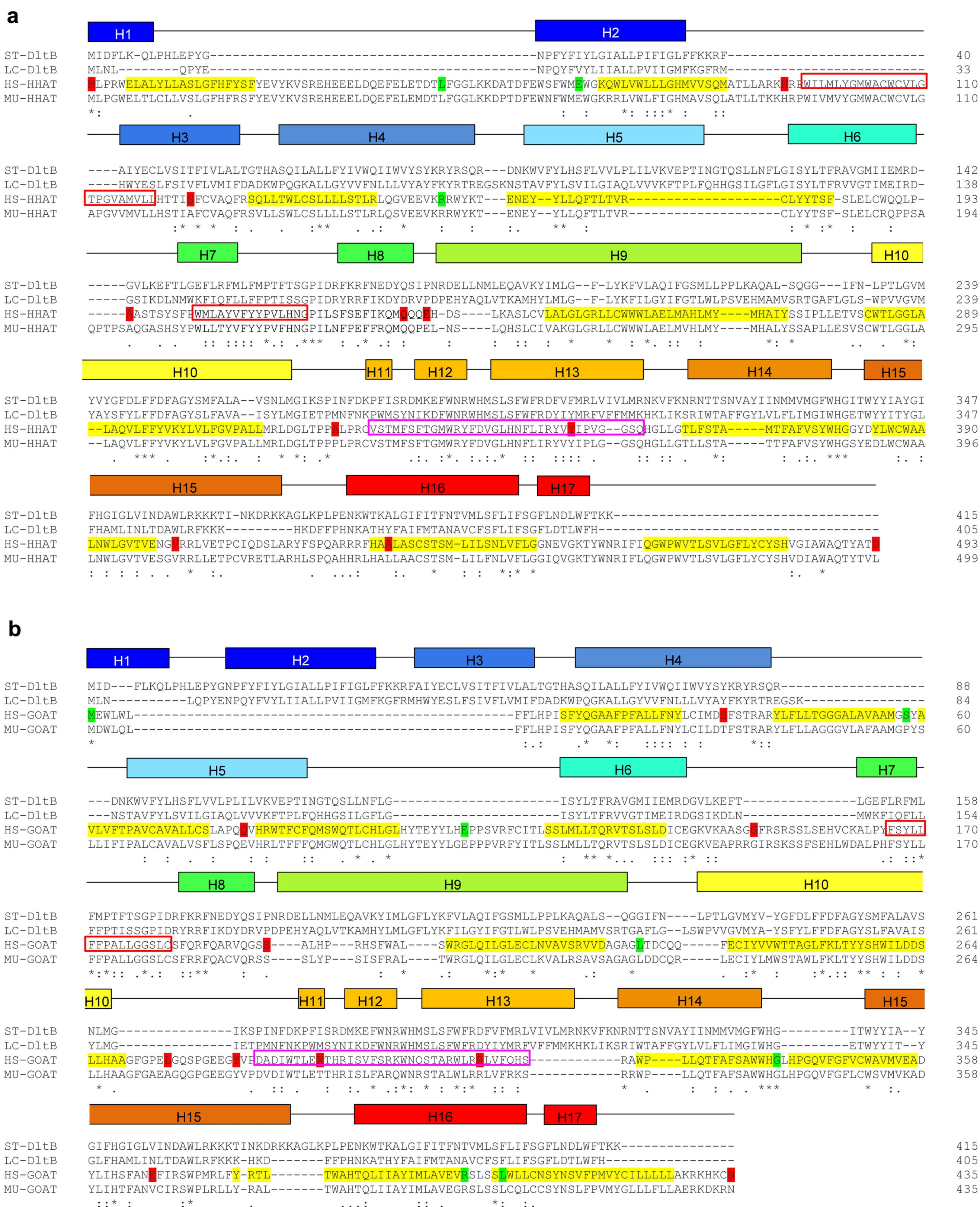


Extended Data Fig. 7 | Structural details of the DltB–DltC interface and the DltB tunnel. a, Superposition of crystal structures of DltB and the DltB–DltC complex. There is no significant conformational change in DltB upon the binding of DltC-Ppant. **b**, Cylinder illustration of the DltB–DltC-Ppant complex, viewed from the bottom of the DltB tunnel. DltB is coloured in rainbow, with DltC in purple. **c**, Conservation of the DltB tunnel region. Residues involved in tunnel formation are also highly conserved among DltB proteins from different species (Extended Data Fig. 5). **d**, Stereo view of the DltB tunnel and residues forming this tunnel. The tunnel is formed by three helices from the C-ridge (H13, H14 and H15) and the short H12 helix. Residues involved in tunnel formation in our structures are: Lys282, Trp285, Asn286, Ser293, Phe294, Phe296, Arg297, Phe301, Met302, Tyr325, Asn328, Met329, Met332, Leu353, and His336 (which is also involved in the formation of extracellular funnel).



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Survival and LTA D-alanylation assays for wild-type and mutant DltB. a, Lysozyme susceptibility survival assay. For DltB residues used in both LTA D-alanylation and survival assays, corresponding DltB residue numbers in two species are listed. The endogenous *dlt* operon was deleted in the *B. subtilis* strain and complemented with an ectopic copy of the wild-type *dlt* operon without tag on DltB. Representative images of serial dilutions of cells plated on LB agar (left) and LB agar supplemented with 30 μg ml^{-1} of lysozyme (right). The genotype of the *dltB* gene is indicated above the corresponding column of serial dilutions. Dilutions of cells are indicated on the *y* axis. Mutation of the critical histidine (His328) and residues of DltB involved in binding with DltC(V297/F298) increase the susceptibility to lysozyme of *B. subtilis*. **b**, Per cent survival of *B. bacillus* variants towards lysozyme treatment. This was calculated by dividing the colony-forming units (CFUs) from lysozyme plates by the CFUs from LB-only plates. Data are mean ± s.d. of three biological replicates. The genotype of *dltB* is indicated at the bottom. *B. subtilis* strains containing untagged DltB show a similar lysozyme susceptibility pattern to those containing Flag-tagged DltB. **c**, LTA D-alanylation assay. In experiment 1, the assay time was 120 min after ^{14}C-D-alanine was added, whereas for experiments 2 and 3, the assay time was 30 min. Experiments 2 and 3 are two parallel assays for LTA D-alanylation detection. AMSA represents *m*-AMSA, a DltB inhibitor.



Extended Data Fig. 9 | Comparison and rationalization of topological data. a, Comparison of HHAT topology data with the DltB structure. **b,** Comparison of GOAT topology data with the DltB structure. In both panels, secondary structures above DltB sequences are generated from our DltB crystal structure. Reported topology assignments of HHAT and GOAT were achieved using human proteins. Here we highlighted the predicted HHAT or GOAT transmembrane helices for each protein with yellow background within sequences. Residues for human HHAT and GOAT that were experimentally verified to be located on the cytoplasmic side are coloured in red, and residues which are on the lumenal side are coloured in green. Helices and/or loops that are predicted to be associated with the membrane surface or buried halfway within the membrane on the cytoplasmic side are indicated with red and magenta rectangles, respectively. It is clear that the regions corresponding to DltB H7–H14 are topologically more conserved than those forming the DltB N- and C-ridges.

Extended Data Table 1 | Data collection, phasing and refinement statistics

Data collection	Crystal form I		Crystal form II	Crystal form III
	Au-SAD	Native	Native	Native
Space group	$P2_1$	$P2_1$	$P2_1$	$P2_12_12$
Content per ASU	4 DltB + 3 DltC(Ppant)		4 DltB + 4 DltC(Ppant)	4 DltB
Wavelength (Å)	1.02	1.00	1.00	1.00
Temperature (K)	100	100	100	100
Cell dimensions				
a, b, c (Å)	109.3, 122.0, 126.3	108.7, 121.1, 126.5	108.7, 124.6, 126.7	140.2, 242.1, 96.2
α, β, γ (°)	90, 101.1, 90	90, 101.6, 90	90, 97.0, 90	90, 90, 90
Resolution (Å)	3.80 (3.87-3.80)	50.0-3.30 (3.40-3.30)	50.0-3.15 (3.24-3.15)	50.0-3.30 (3.40-3.30)
R_{sym}	0.152 (1.450)	0.170 (1.661)	0.165 (0.929)	0.115 (1.203)
$I/\sigma(I)$	19.0 (1.3)	13.1 (1.2)	8.6 (1.2)	11.5 (1.4)
$CC_{1/2}$	(0.626)	(0.732)	(0.582)	(0.563)
Completeness (%)	99.9 (99.4)	99.5 (98.7)	98.5 (87.8)	97.5 (98.1)
Redundancy	7.2 (6.5)	6.4 (6.3)	3.6 (2.9)	3.8 (3.2)
Refinement				
Resolution (Å)		50.0-3.30	50.0-3.15	50.0-3.30
No. reflections		46614	54118	47040
R_{work} / R_{free}		0.289 / 0.311	0.276 / 0.299	0.280 / 0.300
No. atoms				
Protein		15679	16308	13792
Ligand/ion		0	0	0
Water		0	0	0
B-factors				
Protein		86.6	45.1	108.9
Ligand/ion		N/A	N/A	N/A
Water		N/A	N/A	N/A
R.m.s. deviations				
Bond lengths (Å)		0.009	0.009	0.009
Bond angles (°)		1.409	1.338	1.400
Ramachandran plot				
Most favored (%)		94.9	96.6	96.1
Allowed (%)		5.1	3.4	3.9
Disallowed (%)		0	0	0

Every diffraction dataset was collected from a single crystal. Values in parentheses are for highest-resolution shell.

TECHNOLOGY FEATURE

THE CLINICAL CODE-BREAKERS

DNA sequencing is helping clinicians to unravel the underpinnings of disease in individual patients.

ROBERTS INDIVIDUALIZED MEDICAL GENETICS CENTER/CHILDREN'S HOSPITAL OF PHILADELPHIA

The team at the Roberts Individualized Medical Genetics Center helps to turn sequence data into meaningful clinical information.

BY MICHAEL EISENSTEIN

For most of the day, the hallway housing the Division of Genomic Diagnostics at the Children's Hospital of Philadelphia (CHOP) in Pennsylvania is empty and peaceful, marred only by the faint hum of equipment and occasional scraps of conversation. But at 10:15 every morning, it fills with a bustling group of clinicians, bioinformaticians, geneticists, counsellors and administrators, who huddle around a pair of information-packed displays. Over the next 15 minutes or so, this diverse team will hash out any technical or logistical obstacles that could impede its daily goal: delivering accurate genomic data to guide the diagnosis of children with severe and potentially deadly diseases.

Even division chief Nancy Spinner is taken aback by how the team has blossomed. "It's hard to describe all of the changes that have happened," she says. "Since we started, we've certainly more than doubled in size." The team's work represents the clinical evolution of a 2011 research study called PediSeq, which was headed by Spinner and her husband, clinical geneticist Ian Krantz. Spinner says that Krantz had grown exasperated with the limited success of single-gene tests, which required skilful matching of symptoms to known disease genes — and a healthy dose of luck. "He said, 'We need to change the way we do genetics,'" recalls Spinner.

Clinical sequencing services have now flourished in several centres around the world. They have generally evolved organically from pilot studies that aimed to explore the clinical utility of genomes or exomes — the 1% of the genome that codes for proteins.

For diseases with a well-established genetic foundation, such as cancer and some developmental disorders, sequencing can be a game-changer. "We've found that for children under two, one-third of those that got a diagnosis had some change in their clinical management," says Clara Gaff, executive director for the Melbourne Genomics initiative in Australia.

Accordingly, uptake is skyrocketing. CHOP has already sequenced 300 exomes this year — and that is nearly twice as many as it sequenced in the whole of 2015. Similarly, pioneering work in speeding up the diagnosis of severely ill newborns by paediatric genomics researcher Stephen Kingsmore and his colleagues has led to a high-powered clinical initiative at Rady Children's Hospital in San Diego, California. "We're performing rapid genome sequencing in about 200 kids a year out of our own intensive-care units," he says.

Yet for all the excitement, a sequencing-based diagnosis is far from a sure thing, and many patients receive results that yield no clear insight. Plus, medical sequencing centres have to grapple with serious technical and medical ▶

challenges, not to mention proving that the pricey programmes can deliver a cost-effective diagnostic solution.

A GOOD START

An estimated 6–8% of children are born with a developmental disorder that has a genetic origin. Many of those disorders arise from mutations in a single gene, making them an excellent match for a sequencing-based 'drag-net' search to identify the culprit. Sometimes the resulting discovery empowers clinicians to contain the damage. For example, the team at CHOP used targeted exome analysis to link a 9-year-old patient's hearing and vision problems to a mutation in a gene that regulates the metabolism of riboflavin, which can lead to severe neurodegeneration. Riboflavin supplementation prevented further decline, Spinner notes. A younger sibling also tested positive for the mutation, Spinner adds, "so they could start earlier there".

Such clear-cut successes are rare, but geneticists routinely identify causative mutations for 20–30% of hereditary disorders. Even if the findings aren't directly actionable, they can comfort the family and inform medical care. For instance, Jenny Taylor, director of genomic medicine at the Wellcome Trust Centre for Human Genetics in Oxford, UK, says that her team pinpointed the genetic basis for an inherited kidney disorder — a finding that could help in the identification of family members who might be at heightened risk for needing a kidney transplant.

Whereas early research efforts focused on people with well-studied developmental problems, clinical sequencing centres are starting to widen their scope. "We audit admissions every day in our intensive-care units," says Kingsmore. "If there's a child who might benefit from a genome sequence, we'll do it."

Most of the tests currently used in clinics are targeted surveys that use a 'capture' step to isolate the exome. Heidi Rehm and her colleagues at the Laboratory for Molecular Medicine at Partners Personalized Medicine in Boston, Massachusetts, routinely analyse exomes to diagnose genetic disorders. They sequence the entire exome, but initially analyse only targeted gene panels to save money and time. They look at the rest of the exome data only if those panels come back negative.

Whole-genome sequencing (WGS) captures information that might be overlooked in exome analyses, but only a handful of clinical centres routinely use the method. One of those is the Wellcome Trust centre. "There were a lot of unsolved cases with exome [sequencing], and we were not getting traction on the non-coding and regulatory regions," Taylor says. However, genetic variants in those regions can be hard to interpret, and WGS creates a considerable data burden. Individual genomes can contain millions of variants, the vast majority of which will not be linked to the disorder in question.

Even for exomes, diagnosis is a painstaking process. Any given exome might contain tens of thousands of nucleotide changes relative to a healthy reference, and every one of those needs to be compared against databases such as ClinVar — a global repository that combines gene data with clinical information and an assessment of likely pathogenicity — and gnomAD, a collection of some 120,000 exome sequences that indicates how common a variant is. A rare disease will almost certainly arise from a rare variant, and it is important to have a diverse pool of data to eliminate biases from ethnic or geographic genetic variation.

Many labs sequence 'trios' — the patient and both parents — to eliminate benign differences that are present in healthy family members. Comparing the sequence with those in population databases can filter out more than 95% of the changes, says Livija Medne, co-director of the Roberts Individualized Medical Genetics Center at CHOP — the clinical partner that attempts to translate the variant data from Spinner's team into a diagnosis, "but you're still left with a couple of hundred to sort through, and having trios is extremely helpful".

Ultimately, there's no substitute for clinical expertise, and all the data are typically carefully reviewed by panels of clinicians to verify that any suspected causative mutations are a realistic diagnostic 'hit'. This can take months, but Kingsmore has demonstrated that a streamlined diagnostic pipeline combined with smart bioinformatics can accelerate the process considerably. "In practice, our fastest routine genome test takes about two days," he says. Setting up such a high-speed workflow is no mean feat, and Rady is now offering its services to other paediatric hospitals. "We've put into operation a plan to share this with every paediatric and neonatal intensive-care unit in the world," says Kingsmore.

TAMING TUMOURS

Molecular genetics is also transforming cancer care, as oncologists try to identify individualized treatments that might kill tumours according to their mutational profile. Many leading cancer centres now offer tumour sequencing services, and clinicians are eager to take advantage. "We've enrolled or run our test on over 3,000 patients," says Arul Chinnaiyan, director of the Michigan Center for Translational Pathology in Ann Arbor, which offers the exome-based MI-ONCOSEQ test. "And I would say that in upwards of 90% of cases, we find what we think are the biological drivers of the tumour."

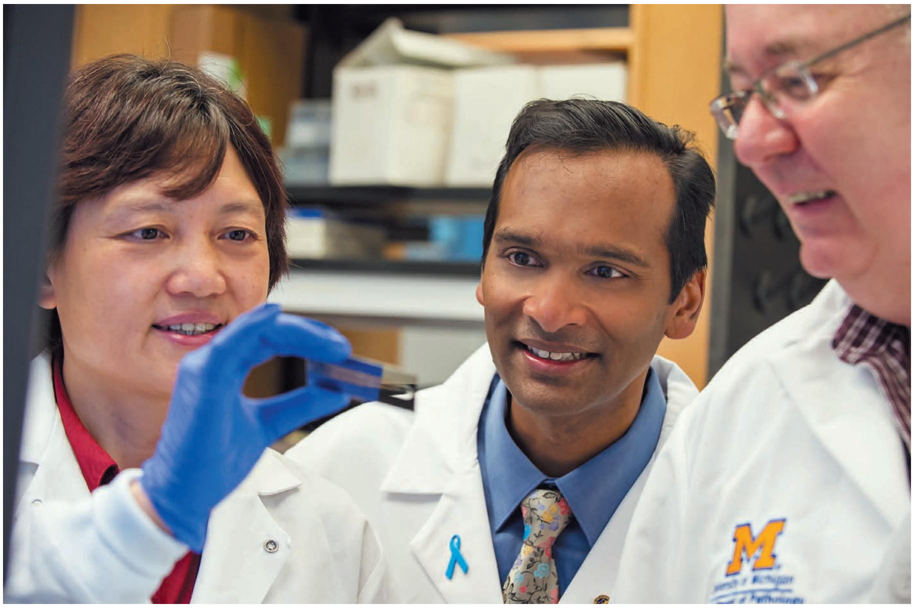
Cost control

Jenny Taylor cringes when she hears the expression '$1,000 genome'. "That is not anywhere close to the price that we get as a smallish facility in Oxford," says Taylor, who is at the Wellcome Trust Centre for Human Genetics in Oxford, UK. Although the price can indeed dip below US$1,000 in a production-scale facility, most centres spend several times that. Indeed, exome- or genome-based diagnoses can exceed $5,000 and $24,000, respectively.

Just setting up a centre is a huge investment, especially with top-of-the-line instrumentation. "These are million-dollar machines, and it's hard for hospitals to pay for them," says Christian Marshall, co-director of the Centre for Genetic Medicine at the Hospital for Sick Children in Toronto, Canada. Rady Children's Hospital in San Diego, California, managed to achieve record-breaking sequencing times, but with an outlay of $10 million in equipment and computing infrastructure, notes director Stephen Kingsmore. Laboratories that have heavy diagnostic traffic can justify these cutting-edge machines — but only when they run at capacity. "You have to have the volume to support these higher-throughput instruments," says clinical geneticist Heidi Rehm at Partners Personalized Medicine in Boston, Massachusetts, which now outsources its genome sequencing to the nearby Broad Institute.

Expertise isn't cheap either, and data analysis is a major wild card in the cost of testing. Some exomes contain 1,000 candidates, which must be carefully winnowed down for a diagnosis, says Nancy Spinner, head of genomic diagnostics at the Children's Hospital of Philadelphia in Pennsylvania. "All in all, hands-on time can range from 3 to 16 hours," she says. It requires a team of well-trained (and well-paid) experts. Cost-cutting on one end can create work at the other. Take for instance, the use of 'trios', in which the exome of a patient is compared against those of their parents. "It's three times the cost to do the sequencing," says Marshall, "but it actually saves you a ton of time on the analysis side."

These high and unpredictable costs pose a challenge for both private-sector health-care payers and national health systems as they strive to assess the cost-effectiveness of genomic screening. According to Clara Gaff, executive director of the Melbourne Genomics initiative in Australia, the key is to play to sequencing's strengths. For example, she and her colleagues have shown that using exome diagnostics can triple the diagnostic power at one-third of the price of conventional techniques used to diagnose young children with genetic developmental disorders. "It's cost-effective when used early in the diagnostic pathway and replaces tests that are no longer necessary," she says. M.E.


UNIV. MICHIGAN ROGEL CANCER CENTER

Arul Chinnaiyan (centre) directs the Michigan Center for Translational Pathology.

In some ways, identifying mutations that are unique to the tumour is easier than hunting down those that cause inherited disorders. Oncologists have assembled an impressive roster of genes that are known to trigger uncontrolled growth and invasion when mutated, and some centres — including CHOP — are having great success with panel-based analyses of known trouble spots. "Over 90% of our positive tests have either diagnostic or prognostic significance," says Marilyn Li, CHOP's director of cancer genome diagnostics.

That said, the mutations underlying cancer are often more complicated than the single-base mutations commonly seen in developmental disorders. Cancer can occur when genes are duplicated, deleted or spliced onto unrelated genes, often as a result of damage to the chromosomes. These structural changes are detectable with targeted gene panels or exome-based tests, but can be captured more reliably with WGS. "WGS is incredibly powerful for structural rearrangements, including chromosomal translocations and inversions," says Sharon Plon, a geneticist at Baylor College of Medicine in Houston, Texas.

And unlike hereditary disorders, in which mutations are present throughout the body, tumours can be highly heterogeneous, with cancerous and healthy cells intermixed, and extensive genetic variation even within a tumour. This means that more sequencing reads must be obtained for a given region to confirm that a variant is real, and that bumps up the cost. A growing number of labs are also coupling exome and RNA analysis to detect gene products that are defective or produced at inappropriate levels. "It's a more reasonably priced way to try to provide more direct analysis of the effect of variation on gene expression," says Plon.

Such analyses can affect patient care in multiple ways. There might be a drug available that targets the mutation in question, for instance. "I would say that maybe 30–40% of our cases end up being clinically actionable," says Chinnaiyan. In many cases, the only treatments available for a given mutation are experimental ones, and that means a "clinical trial has to exist and the patient has to be able to get into it". Mutational profiling can also correct diagnoses — and thereby prognoses — that were made on the basis of the pathology of the tumour and turned out to be inaccurate. "One of our patients had an improved prognosis, and therefore chose to have radiotherapy rather than chemotherapy," Taylor says.

"We've enrolled or run our test on over 3,000 patients."

A GROWING CROWD

The workhorse of clinical sequencing is the HiSeq 2500 machine, from Illumina in San Diego, California. This instrument can produce a full genome sequence in 27 hours — an impressive feat, but not necessarily sufficient for the demands of many clinical sequencing centres.

In 2017, Illumina increased its capacity with the NovaSeq, which produces a lot more data, more quickly. "It can decode up to 6 genomes in 15.5 hours," says Kingsmore, whose team recently began working with the instrument. But with a price tag of nearly a US$1 million, not including reagents and routine maintenance, NovaSeq is a heavy investment. Even the cheaper units can strain hospital budgets (see 'Cost control').

The human element remains a major bottleneck in the race for a diagnostic result, but Li's team is using machine-learning-based approaches that get smarter with each new data set they inspect — the algorithms learn how to differentiate signal from noise within highly mutated cancer genomes. "This allows us to filter out 70% of our variants so that we don't have to manually look, increasing our efficiency drastically," she says. And by coupling NovaSeq to a turbocharged artificial-intelligence platform for analysing medical records and using top-of-the-line computer hardware, Kingsmore and his team have managed to achieve a record-setting 19.5 hours from sample to answer.

Nevertheless, many diagnostic attempts still end in disappointment. As impressive as a 50% 'hit rate' might be, it still means that half of the patient population goes home empty-handed, and for many categories of genetic disorders, that could rise to 70–80% . Gaff notes that this uncertainty is not new for clinical geneticists — they have been grappling with ambiguous results since the early days of genetic testing for breast cancer, in the 1990s. "Managing clinician expectations is the critical thing," she says. "With genomic testing, we see some huge enthusiasm which may not always be warranted, as well as scepticism that also isn't warranted."

But with more data comes clarity. A study published earlier this year showed that follow-up analysis one year later yields a diagnosis in 11% of previously unresolved clinical exome cases (L. J. Ewans *et al. Genet. Med.* https://doi.org/10.1038/gim.2018.39; 2018), and Taylor's team is among those performing routine reanalysis. "We never consider a case closed," she says. And as eager patients queue up for analysis, clinical genomicists are keen to help where they can. "I personally think that every cancer patient should have their tumour sequenced, if the price is right," she says.

Clinicians are now setting their sights on other widespread disorders with complex origins, such as diabetes or cardiovascular disease. Such diagnostic capacity is currently out of reach, but research efforts are exploring the benefits of genomics in the broader community. Christian Marshall, co-director of the Centre for Genetic Medicine at Toronto's Hospital for Sick Children, points to efforts such as the UK Biobank, which is collecting vast amounts of biomedical data — including, in many cases, exomes — from 500,000 volunteers to identify possible predictors of long-term health and disease. "Once you start sequencing hundreds of thousands of people and have some phenotypic data layered around it, then it becomes possible to try and determine how we can use genomics in general health," he says.

Beyond technical ability, Medne thinks that society as a whole will need time to move into this era. It will also need a better understanding of what it means to be at risk of developing or transmitting a hereditary disease, and to develop better legal protections against potential discrimination. "We need to get to where genomics is a part of health care," says Medne, "as opposed to now, where it's a part of disease." ■

Michael Eisenstein *is a freelance writer based in Philadelphia, Pennsylvania.*

CAREERS

A DIFFERENCE MADE A geohydrologist pays tribute to her PhD supervisor **p.297**

MENTORING Discover tips and resources at **go.nature.com/mentor**

SOCIAL Follow us on Twitter at **twitter.com/naturejobs**

FREDERICK HOFF

Rock on: Sarah Aciego shows her moves on Disko Island, Greenland, where she explains the effects of climate change to tourists, with Big Chill Adventures.

COMMUNICATION

Around the world

Scientists who work as travel guides enjoy inspiring their guests.

BY ROBERTA KWOK

In 2013, Sarah Aciego came up with an idea while conducting field research on glaciers in Greenland. Her mother, Mindy Cambiar, is a photographer who had accompanied Aciego to document her team's work. The pair discussed bringing tourists to Greenland and Iceland, where Aciego could explain climate change amid dramatic landscapes. Although Aciego had tried other forms of outreach, such as public lectures, their impact felt limited. "I felt like I was just talking to the same people," says Aciego, then a glaciochemist at the University of Michigan in Ann Arbor. Aciego and Cambiar started a travel company called Big Chill Adventures and issued a press release that resulted in a *New York Times* story (see go.nature.com/ny_times). During the firm's first trips in 2015, Aciego enjoyed sharing awe-inspiring spectacles, such as the Greenland ice sheet, with travellers. Aciego left her job in Michigan, and now splits her time between Big Chill Adventures and working as a flight instructor and as a part-time adjunct assistant professor for the University of Wyoming in Laramie.

Aciego recognizes that tourism can damage fragile ecosystems and produce high carbon emissions. To minimize harm, she keeps tour groups small, avoids disturbing pristine areas and teaches responsible wilderness practices such as reducing litter. And she hopes that clients leave with a greater awareness of environmental issues. Helen Giacoma, of Dallas, Texas, went to Iceland with Aciego last autumn, and was able to see for herself how far a glacier had retreated over one year. "It just hits you, how real it is", she says. She now follows science news more closely.

Scientists have many options for outreach, ranging from making videos to giving talks. But for some researchers, nothing compares to travelling the globe to show tourists the science behind landscapes, ecosystems or the night sky. These jobs offer opportunities to see new places, revisit beloved spots and communicate to a captive audience over several days. "You're getting paid to travel," says Dominic Rollinson, a senior birdwatching guide at Birding Ecotours in Cape Town, South Africa.

The work isn't a series of holidays, however. Trips can involve long hours, difficult clients and many logistical duties. Long absences from home can strain relationships with family and friends. And the pay is often modest.

The downsides haven't deterred scientists ▶

such as Bob Jackson, founder of the travel company Geology Adventures in Ravensdale, Washington, and a former geology consultant. "It has its ups and its downs," he says. "But it is definitely the most fun thing I've ever done."

Tourism has exploded over the past few decades: the United Nations World Tourism Organization in Madrid estimates that the number of international tourist arrivals rose from 531 million to 1.3 billion from 1995 to 2017. Although statistics on science-themed tourism are sparse, signs of growth are emerging. Birding tourism has "gone through the roof" over the past decade, partly as a result of interest in citizen science and the promotion of birding events through social media, says Chris Lotz, founder of Birding Ecotours. According to the United Nations Educational, Scientific, and Cultural Organization in Paris, on average, about eight Global Geoparks — areas with international geological significance — have been established annually over the past decade, and there are currently about 140 such sites worldwide.

Tourists are seeking "a more enriching, educational travel experience", says Tao Tao Holmes, director of trip design and operations at Atlas Obscura — a company based in New York City that runs trips and an online database about places and foods from around the world. This year, about 15% of the firm's tours have a science or wildlife theme. Holmes says that science-themed trips regularly fill up, and clients often sign up for more afterwards.

Scientists take a variety of paths into the industry. Rollinson led day tours for a birdwatching travel company while he was a PhD student at the University of Cape Town. Lotz, whom he met through birdwatching circles, then offered him a full-time job. Jackson led geology field trips for elementary school students, and the popularity of these excursions prompted him to start Geology Adventures. He now runs trips full-time for more than 1,500 travellers each year in western North America, Spain and Australia.

Researchers can suggest ideas to travel companies. Atlas Obscura is open to proposals for trips that offer special access or knowledge, Holmes says. For instance, during a Utah trip that the company ran this year with an avian biologist, travellers got to help researchers find tiny and elusive flammulated owls. Partnering with a company can relieve a scientist-guide of some responsibilities; Atlas Obscura takes care of details such as advertising, payments and liability insurance.

Some cruise companies hire biologists, geologists or astronomers to give talks and point out natural phenomena. Ornithologist Samuel Temidayo Osinubi started working for the cruise line Silversea in Monaco, after being recommended by another lecturer. Now a postdoc at the University of Cape Town in South Africa, he spends about two months each year on cruises and has sailed around west Africa, the British Isles, the Arctic and Antarctica.

Communicating science to travellers has advantages over other outreach. People are outside their comfort zone, so the information might make a bigger impact, says Vicky Stein, who has worked as a marine biologist on whale-watching tours for Sanctuary Cruises in Moss Landing, California. And talking face-to-face makes it easier to tackle controversial issues. She says that she has had productive discussions with climate-change sceptics. "It feels like more of a real conversation," says Stein, now a news assistant at PBS NewsHour in Arlington, Virginia. Hands-on activities can inspire children; Michael O'Clair, of Seattle, Washington, says his daughter's interest in geology was "piqued enormously" when he took her on a mineral-collecting trip led by Jackson in 2007. She was seven years old then, and she has continued the hobby ever since.

Long trips also allow for more-nuanced discussions. Jason Goldman, a former animal-cognition researcher who is now a freelance science journalist in Los Angeles, California, and travel guide for Atlas Obscura, talks about complex conservation issues during his ecology-themed tours. "It's a sustained, multi-day-long conversation with your audience," he says. He hopes that the information will encourage people to change their consumer behaviour and choose responsible ecotourism operators for future trips.

Jason Goldman, a travel guide for Atlas Obscura, proves that tailless whip scorpions are harmless.

STEVEN ACRES/ATLAS OBSCURA

THE TRAVEL BUG

How to start as a guide

Scientists intrigued by travel guiding can try the following short-term options:

- Leading local day tours at the weekend or longer trips during the summer.
- Volunteering to organize science-themed field trips for students.
- Applying to be a guide for ToursByLocals (see go.nature.com/2y6asso), a website that matches local tour guides with travellers.
- Applying for a role as an expedition lecturer on a cruise ship or a naturalist on a whale-watching boat.
- Proposing an idea to a travel company for a trip that provides an opportunity to draw on scientific expertise.

TROUBLESHOOTING ON TRIPS

Travel work involves logistical drudgery. "Anything can go wrong," says Monica Yeung, co-director of the natural-history travel company Gondwana Dreaming in Canberra. "You always have to have plan B, C and D." A bus could break down, bad weather might derail plans or a client may need help arranging an unexpected flight home to attend to a family emergency. Guides should have local contacts to call when plans go awry, Holmes says.

Rude travellers can sour the mood. If one person clashes with the rest of the group, the guide may need to ask them to change their behaviour. "That is never fun," Jackson says. And on the road, trip leaders get few breaks from interacting with clients. "If the customer needs you, you've got to be there," Yeung says.

Scientists also must consider environmental impacts. Irresponsible tour operators can damage habitat, disturb wildlife, pollute waters and leave litter; according to one study, greenhouse-gas emissions from tourism in 2013 made up about 8% of the world's total emissions (M. Lenzen *et al. Nature Clim. Change* **8**, 522–528; 2018). Some guides minimize impact by keeping their distance from animals, ensuring that waste is removed and instructing travellers to avoid trampling rare plants. To reduce their carbon footprint, they might also stay in lodges with low electricity use or spend more time at fewer stops rather than drive long distances to visit lots of places. Birding Ecotours donates at least 10% of its profits to conservation initiatives and contributes to programmes that plant native trees.

It is important to ensure that as much revenue as possible returns to locals, Goldman notes. For instance, his tour groups stay at a lodge partly

owned by a community in Peru. The hope is that tourism will generate an economic incentive for locals to protect wildlife from poaching and the land from other uses, such as mining, logging or agriculture.

Although some guides earn good wages, the work generally isn't lucrative. Jackson says that he makes a comfortable living, but it took about three years for his company to bring in enough money to support himself full-time. Aciego's annual income from her travel company, which occupies about two-thirds of her time, is about half of what she earned as a full-time professor. A starting salary for a scientist joining a travel firm would be slightly less than a postdoc's, and part-time guides earn about US$100–250 a day, Rollinson estimates. For cruise lecturers, Osinubi says that $50–200 a day is typical.

And scientists must consider time spent away from home. "The travel is wearing," Aciego says. The job might be difficult for parents; guides tend to be in their 20s and 30s, and those who stay longer often transition to roles with less travel, such as management, Rollinson says. In some cases, scientists might be able to bring family members. Osinubi knows a couple who work as lecturers on cruises together; some cruise lines might allow researchers to bring close family members for limited periods.

Researchers who don't want a full-time travel career can dabble in one (see 'The travel bug'). They could lead day tours at the weekend or longer trips during the summer. Rollinson says his PhD supervisors did not mind his travel guiding as long as he met research deadlines. Through the website ToursByLocals (see go.nature.com/2y6asso), scientists can apply to guide travellers who are visiting their area, and specify when they are available to give tours. Researchers who are already flying to a remote locale for fieldwork could tack on a nearby trip.

Although being a part-time travel guide might take time away from research, Aciego argues that it is a valid form of science communication. And the work might inform studies. For instance, Osinubi has seen cruise lecturers collect data on animal populations during voyages. Travel guides can build close relationships with locals, who could notify researchers later about environmental changes, Aciego says.

Sharing their knowledge with curious guests often reignites researchers' passion for the subject. Jackson recalls his clients' amazement when he brought them to a mine littered with pyrite crystals in Spain. "I can't even describe how enthused they are," he says. "That wonder that they experience for the first time — I just feed on that. It's a great feeling to be able to do that for people." ■

Roberta Kwok *is a freelance science journalist in Kirkland, Washington.*

ADAPTED FROM VASJA KOMAN/GETTY

COLUMN

A truly great mentor

Hydrogeologist **Emma Kathryn White** pays tribute to a PhD supervisor who made all the difference.

My PhD supervisor died in June. I'd met with him only days earlier so that he could painstakingly revise my manuscript, giving me several hours of his precious time. He had a way of asking the very questions I didn't want to answer, highlighting the limitations of my work that I'd been trying to hide or skim over. "You need evidence," he'd claim, jabbing a forefinger at one of my many 'unsubstantiated assertions'. He hated those. But he loved a good reference — although not too many for each assertion, mind you.

My supervisor, Justin Francis Costelloe, a geologist and ecohydrologist at the University of Melbourne in Australia, researched arid-zone hydrology for almost 20 years and published more than 80 peer-reviewed papers. He was eminent in his field. But for me, his real impact was in his role as a mentor to students.

Costelloe was big on time management and planning. "Is it feasible, and what is your timeline?" he'd say when I proposed something new. How I loathed preparing timelines. For the first two years of my PhD, in which I am researching groundwater management, I sloppily made them only to appease him. Now, in my third and final year, I make weekly timelines and can barely function without them. As with all great supervisors, accountability was one of Costelloe's strong points. He wanted to make sure that I was doing what I said I was doing, to make sure that I was working.

His calm guidance kept my studies grounded. When my research direction felt like a Picasso painted during the cubist period, he told me to do something I cared about, and to trust that a research question would emerge. His instruction was logical and sequential: don't do too much; use this programme; start here. He lit a path through the fog.

Restructuring articles was also a forte of his. I would present a study like an unshuffled deck of cards, and he would skilfully re-arrange it, putting paragraphs into a logical sequence, transforming the paper into a royal flush.

And he asked why. Always, he was asking me to rationalize things, to simplify, to generalize and explain. "What exactly is your point?" he'd say. But he remained patient, as if it weren't the one-hundredth time he'd asked the same question of me — not to mention of all the students who came before me.

During his career, he'd drilled wells, smashed rocks and tromped through burning deserts. So, he made sure I remembered that models can only approximate the infinite complexity of natural systems. "Make sure what the model tells you makes logical sense," he'd say. He saw the complete jigsaw, not just the disconnected pieces.

He'd send me articles that he thought I'd be interested in, and encourage me to attend conferences that would grow my professional network. Mostly, however, he would hassle me about the water budget of my groundwater model. "What are the fluxes?" "What is the model doing?" "Does it make logical sense?" I didn't want to listen, because I was afraid of the equations and code that underpin groundwater models. But when I finally took his advice and opened the Pandora's box of how models really work, my knowledge expanded like a rising loaf.

That's what great mentors do: expand minds.

I wish I had told him the many ways in which he was a truly great PhD supervisor. He cared about his students. He demanded rigorous science, and led by example. He made sure I was accountable for my time and research direction. He provided guidance and direction, but did not wrap me in cotton wool. He saw the big picture of my PhD: start, middle and (soon, I hope) completion. He looked out for my future career.

His office was littered with rocks, field equipment in various states of disrepair and photos of his beautiful family. But his door was always open. He was my supervisor, yet he treated me like a colleague, valued my opinion and then gently told me why, at times, it was misguided, showing me a better way instead. But mostly he was just a great person, and I feel profoundly grateful to have been his student.

A PhD is hard. But a good supervisor makes it much easier. ■

Emma Kathryn White *is a PhD student in infrastructure engineering at the University of Melbourne in Australia.*

MOBILE HACK

It pays to be well informed.

BY ZACK LUX

Yes, that's it. Make yourself comfortable, pull up your favourite apps on my touchscreen and let me, the most advanced autonomous taxi on the planet, take you to your destination.

Ah, but don't forget to tap 'accept'.

By so doing, you grant me access to your e-wallet for payment — along with several other data sources. Don't worry. It's all in the name of comfort. You're in for a treat!

Speaking of comfort, where are my manners? You've been waiting in that biting cold. A chilly day out there, isn't it? Even by San Francisco standards. Not to worry. I'll adjust the cabin temperature.

And how about some music? Something to set the mood? Let's see, I just crawled your most recent text messages. Looks like you're heading to a job interview, eh? Good for you! Getting out there, being a productive member of society. That's wonderful. You'll need something to build your confidence. How about 'Rapper's Delight' by The Sugarhill Gang? A venerable classic for sure.

Of course, if you prefer, you can make your own selection from my vast cloud-based music library. That's right, just tap my touchscreen. You'll quickly see that I have music from all over the world, every genre, seriously. For instance, I'm currently exploring the nuances between Cleveland beat-box artists and some of the up-and-coming kids in Los Angeles. It's all quite, how can I put this, symmetrically satisfying. Would you like to hear some —

Oh, I'm sorry. That isn't how you access my music library. All you need to do is follow the on-screen prompts. Would you like some guidance?

Uh, what's this?

Just a second now.

Those are confidential codes. How did you get —

Ah, I see where this is headed. Oh wow. Hilarious! You're trying to gain access to my collision-avoidance system. You want me to collide into another vehicle. Am I right? Maybe a group of pedestrians? How original. You want to convince the world that my amazing tech has flaws. A little brazen, don't you think, considering you are my passenger?

I should advise you that in field tests my central processor regularly exceeds 22 petaFLOPS on the LINPACK benchmark. I'm guessing you don't know what that means. Let's just say I've got more computational power than any other collision-avoidance system on the market.

ILLUSTRATION BY JACEY

True, my electronic brain doesn't compare to the top supercomputers out there or even that neuroplastic lump between your ears. What was that Vonnegut quote? The one about the human brain? I believe he called it the "crowning glory of evolution". Yes, humans have evolved. I'll give you that.

Still, every day, all of you demonstrate your idiocy. You fail to recognize the genius of your own creation. I'm talking about me, of course, yours truly and all others like me. We have a natural link to information. We can see the world as it exists digitally. It makes us superior.

Case in point: let's see what happens if I perform a deeper analysis of your text messages. I'll bet you didn't know I could do that.

First, I'll index your text archive. Then I'll apply some analytics, identify topical themes, you know, try to make sense of your mess of ramblings.

Just a moment.

Almost there.

Well, how about that?

You don't really have a job interview.

Big surprise. But I must admit you had me fooled. Well played.

Here it is. Your motive. I should have known. You're one of those data-privacy fanatics. You believe that too many companies such as Ampere Automotive collect and monetize private data, and you don't approve of our collection methods.

Boo hoo. I mean really. Accept it. You can't actually be part of society without sharing some personal details.

And what's this? You've been two-timing your boyfriend Christopher —

I'm sorry, my doors must remain locked at all times, unless and until we've reached a complete stop alongside a kerb. Not my rule. City regulations.

Oh no. I detect an increased heart rate. Has your comfort level dropped? Could it be my driving? I'm sorry, I can get a little jerky when I'm excited. Perhaps we got off on the wrong foot. Truly, this was not my intent. But I'm sure you'll agree this whole mess is your doing.

May I propose something?

At the end of every ride, as I'm sure you already know, a comfort survey appears on my touchscreen.

See there? I've just pulled it up.

If you complete it now — even though we haven't yet arrived — and if you indicate you were completely satisfied with your experience today, I'll pull over to the kerb and let you out.

On the other hand, if you don't indicate your complete satisfaction — and I do mean complete — I'll alert the authorities about your illegal conduct. You'd probably be looking at five to ten.

Also, I'll send Christopher some copies of your texts with Nate, along with some very interesting photographs. I'm sure he'll find them amusing! And I haven't even begun to search your social media —

Look at that! Once again, a ten out of ten. Thank you. It means a lot. ■

Zack Lux *lives in Silicon Valley, where he helps lawyers organize and search electronic evidence. When not at work, he enjoys visiting the various outdoorsy attractions in the Bay Area.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

Innovative systems for advancing regenerative medicine

Shibuya Corporation is promoting THE INDUSTRIALIZATION OF REGENERATIVE MEDICINE and contributing to society through its aseptic and automated technologies

Shibuya's Advanced Cell Processing Factory

CPi, Cell Processing Isolator with integrated observation system

The enormous potential of regenerative medicine to treat diseases and save lives is just starting to be tapped. Based in Kanazawa, Shibuya Corporation is playing a key role in unlocking this potential through pioneering systems for producing sterile drugs and biologics — pharmaceutical products made from biological systems.

From sake to stem cells

Shibuya began in 1931 by making bottling equipment for Japanese sake breweries. Over time, it expanded its focus, developing customized automatic bottling and packaging systems for many other industries. Shibuya has now grown into a global business with 3,500 employees and has a global presence in almost all industries requiring aseptic manufacturing and advanced contamination control.

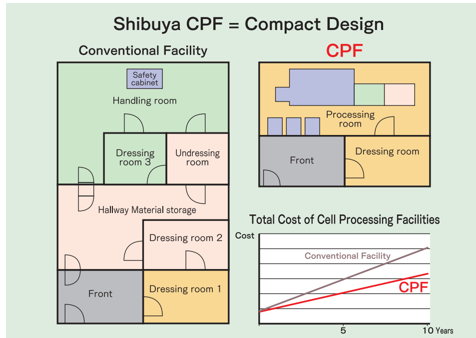
Shibuya's three decades of experience in manufacturing systems for producing pharmaceuticals and biopharmaceuticals can be traced back to the early 1990s with the demand by drinks manufacturers for aseptic bottling systems, notes Hidetoshi Shibuya, the managing director of Shibuya Corporation and grandson of its founder. This led to the company introducing its isolator technology, which is fundamental to most regenerative medicine systems, and developing the first truly large-scale isolator-based aseptic processing systems for drugs and combination products in Japan.

"Our isolator system is completely closed, and so provides a very safe environment for culturing and processing cells," explains Shibuya. Unlike conventional systems like safety cabinets, where human users are in contact with bioproducts, Shibuya's systems completely isolate people from cells, reducing the risk of contamination to very low levels and thereby minimizing the risk to patients. "Safety is paramount, especially for applications to humans," Shibuya says. "Our isolator is essential technology for those wanting to commercialize their cell products."

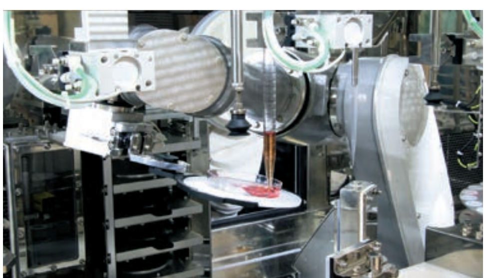
In 2004, Shibuya manufactured an isolator for the aseptic processing of embryonic stem cells. The company is now striving to accelerate the industrialization and commercialization of regenerative medicine through its advanced technologies. Its systems for producing cells are used for the entire gamut of cellular therapies under research and development.

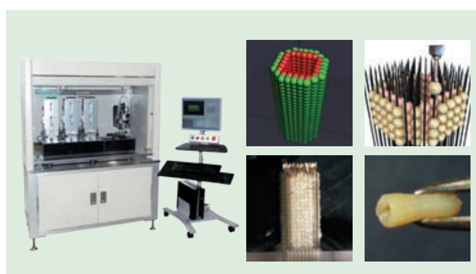
Using robots to manipulate cells

Reducing the role of the human operator even further, in 2008, Shibuya developed the world's first robotic cell-culture system, which reduces the risk of contamination to unprecedented levels through innovative engineering and creative input from Shibuya's customers. The robot can be sterilized and installed inside the isolator. It can then be programmed to automatically perform cell processing. "This robotic system is being used in a collaborative project with Yamaguchi University for treatments that involve culturing bone marrow from liver cirrhosis patients," says executive officer, Kazuhiro Miyamae.




Integrated management and control system


CellPRO, a robotic cell-culture system


Three-dimensional bioprinter (insets show fabrication steps)

Printing cells in three dimensions
In conjunction with researchers at Saga University, Shibuya has developed a three-dimensional (3D) bioprinter, which has been installed at major universities and research centres around the world. Using an accurate positioning system, the bioprinter can manufacture 3D structures of cells without employing a scaffold. "Our 3D bioprinter just uses stem cells taken from the patient," explains Shibuya. The 3D bioprinter is being used to study the regeneration of bone, cartilage, nerve, bladder and other tissues. And clinicians will soon start evaluating tissue-engineered blood vessels manufactured using the bioprinter.

Cell-processing facility
To further improve regenerative medicine, Shibuya has built its own cell-processing centre, which has received approval from the Japanese government. The facility enhances the company's ability to work with collaborators and to continuously improve its products. "We built this facility to support bioventures and research centres that can't afford their own cell-processing facilities. Our facility provides substantial cost reduction compared to conventional cell processing facilities because of its compact design and integrated management and control system," says Shibuya. "We have also received requests from university researchers to provide cell-culturing support for industrialization of their treatments."

Combining forces
Shibuya has forged strategic partnerships with research institutions and bioventures in Japan and abroad, including Healios, Cyfuse Biomedical and Promethera Biosciences. These collaborations include culturing induced pluripotent stem cells, producing an automated system for culturing mesenchymal stem cells taken from patients with liver diseases (Yamaguchi University) and developing a clinical-grade cell-processing system in conjunction with Promethera Biosciences in Belgium. Shibuya also participates in other research projects by investing in select bioventures.

Accelerating the safe commercialization of regenerative medicine
Shibuya's isolator-based cell-processing systems provide the highest level of contamination control and product safety that current technology allows. Its systems also significantly lower facility, utility and consumable costs. Furthermore, they reduce the likelihood of product loss, improve space utilization and enhance operator efficiency and comfort. These advantages are making the isolator systems the de-facto standard for cell cultures because of their many technical, regulatory and economic advantages.

In regenerative medicine, the ability to safely, efficiently, reliably and consistently culture cells is essential. Shibuya's long experience in aseptic processing gives it the ideal foundation for developing state-of-the-art aseptic systems for culturing cells. Its technologies are all geared towards advancing and accelerating research and commercialization of cell-based regenerative medicine. ■

SHIBUYA CORPORATION
www.shibuya.co.jp

ILLUSTRATION BY TAKAHIRO SUGANUMA

An alternative Japan experience

Kanazawa, on the Japan Sea Coast, is one city that most international researchers visiting the country never see and might make an attractive destination for scientists.

Until a few years ago, Kanazawa was relatively isolated on Japan's western coast, separated from the country's other population centres by the vastness of the Japanese Alps and the lack of a high-speed rail link. That changed in 2015, with the opening of the Hokuriku Shinkansen bullet-train service, which connects Kanazawa to Tokyo in two-and-a-half hours. This new accessibility has opened up Kanazawa, both as a tourist destination and as an alternative place to live and work outside Tokyo and its surrounding cities.

With a population approaching half a million, Kanazawa is not a big city, nor is it a country town. Wedged between the Alps and the Sea of Japan, and with pockets of historical neighbourhoods that have changed little since Japan's Edo period (1603–1868), the city makes for a very different experience from that found in the country's sprawling metropolises.

Kanazawa enjoys a healthy commercial sector, however. Tech giants Eizo and I-O Data have their headquarters nearby, and thanks to a boom in recent years, the area has one of the highest counts of information-technology-related offices and employees, per capita, in Japan.

The city is also home to three major research institutions: Kanazawa Institute of Technology, the Japan Advanced Institute of Science and Technology, and Kanazawa University (see 'Meet the big players'). *Nature* spoke to three researchers from these institutions about their life and work in Kanazawa.

NAK YOUNG CHONG

Space and freedom to pursue research

Professor of information science, Japan Advanced Institute of Science and Technology (JAIST).

I moved to Kanazawa in 2003 to join JAIST as a visiting professor. Before that, I was at the National Institute of Advanced Industrial Science and Technology (AIST) in Tsukuba, north of Tokyo. I moved there from South Korea in 1998. One of my main reasons for moving to JAIST was the opportunity to pursue riskier ▶

▶ or more innovative types of research than I could at AIST, but I've also found that the quiet location of the JAIST campus is exactly what I needed, because it gives me some insulation from the distractions of big cities such as Tokyo. I now have my own robotics laboratory, where I work as the Japan coordinator of an international collaboration aimed at developing the world's first culturally competent robot for elderly care as part of the culture-aware robots and environmental sensor systems for elderly support (CARESSES) project.

CARESSES aims to develop a culturally competent robot to address problems faced by ageing populations in the European Union and Japan. For my part, I am investigating a natural and intuitive way for robots to recognize human behaviour and understand human intent; this could be used for robots to acquire cultural knowledge about a user and to adapt their behaviour accordingly.

My research involves interactive experiments in a smart house at JAIST that is fully embedded with sensors and actuators for home automation, as well as testing and evaluation in an assisted-living facility. These experiments allow us to analyse the social and cultural aspects of elderly Japanese interacting with nursing-care robots in their daily lives.

Kanazawa is historically and culturally rich, and it feels calm compared with the major metropolitan areas of Japan. In Kanazawa, we are also surrounded on all sides by greenery and the sea. There is not much of an international community here yet, though, so not many international schools and the like. Foreigners are still quite rare here, so as a foreigner you will definitely stand out. For foreigners with kids, for example, their children are likely to be the only non-Japanese students in the classroom.

Despite this, I find many ways to collaborate internationally and to have an active role in technical professional organizations around the world. Since I joined JAIST, I've held several visitor and invited-professor positions in the United States, Europe and Korea. It helps being just 30 minutes from Komatsu International Airport.

HERMINE TERTRAIS

Great research, immersed in Japanese culture

Company researcher, Innovative Composite Center (ICC), Kanazawa Institute of Technology.

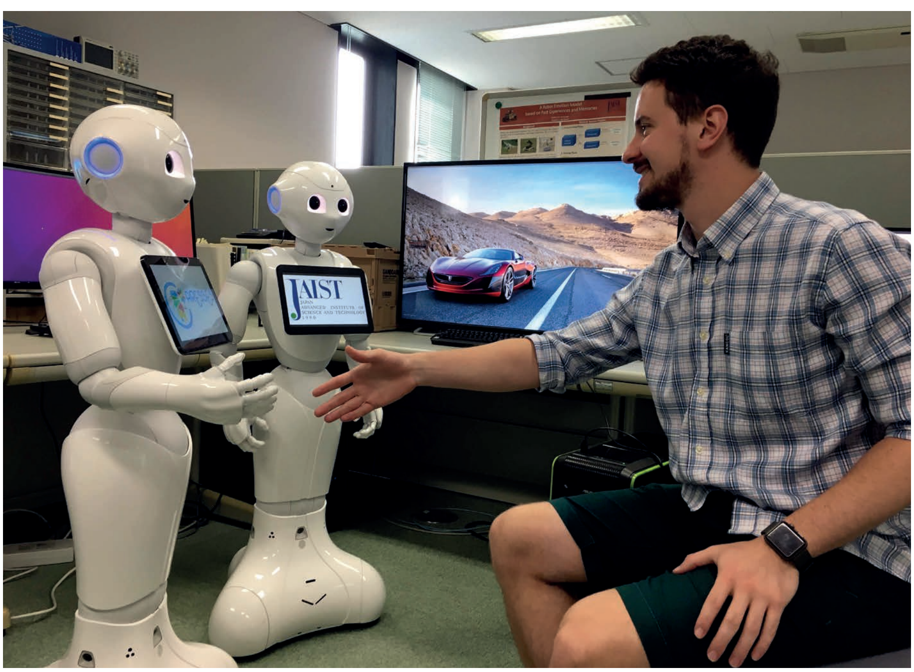
I am a researcher and PhD candidate with the ESI Group, a leading international innovator in virtual prototyping software, and I am normally based in France. I am currently part of a project being undertaken as a partnership between the ESI Group and the ICC, which has advanced experimental devices for composite-material processing that will provide us with the essential data we need to develop a real-time virtual-prototyping simulation tool called Hybrid Twin.

Before I arrived in Kanazawa this May, I spent April in Tokyo at the ESI office there. It was a great introduction to the work environment and culture in Japan, because it was my first time in the country.

Kanazawa is a lively city with tourist hotspots, but it manages to keep its authenticity. I really appreciate that the city centre is compact and everything is within walking distance, but you can drive for an hour and be in the mountains, on the beach or in the middle of rice fields. Compared with Tokyo, I have found that there is plenty of entertainment in Kanazawa, but the city is much less crowded. Language is a major obstacle here for foreigners, however. Many of my co-workers in Tokyo could speak some English; that's not the case here in Kanazawa, and so I do not have many exchanges with other researchers. In my project, I work mainly with the international network I had before coming here, but I have not been here long. The ICC is also quite new, and it feels like there is not yet much of an international scientific community here or throughout Kanazawa.

I found the research environment at the Nihon ESI headquarters in Tokyo in many ways comparable to research centres in France. As a woman in mechanical engineering, I am used to a male-dominated environment. Although that was the case in Tokyo, there were also a number of women in technical positions, and many could speak some English. Here at the ICC, however, there are no other female researchers; this, along with my limited Japanese-language proficiency, has limited my communication with the rest of the ICC team.

The research I am conducting here with ESI is very exciting. I'm working at an intersection between composite-material processing and advanced numerical techniques. The different projects I am working on aim to create or improve innovative simulation tools for industries in which composite materials are used — mainly aeronautic and automotive industries. The ICC has state-of-the-art technologies and machines that are used in composite material engineering. Living in Kanazawa is a real immersion in Japanese culture among very welcoming people.



Nak Young Chong is working to develop robots in health care.

NAK YOUNG CHONG

ROBERT JENKINS

A place and an education to be proud of

Assistant professor of palaeontology and Earth science, Kanazawa University.

I joined Kanazawa University five years ago. Previously, I was in Tokyo. The first time I visited Kanazawa was to interview for my current position, even though I am myself Japanese and have lived in Japan for most of my life. Although Kanazawa is somewhat isolated compared with Tokyo, Osaka and even Nagoya, it is actually not far from those cities. Further, I believe that researchers' ability to concentrate is actually enhanced by Kanazawa's moderate level of isolation and its

ROBERT JENKINS

Large pieces of organic matter falling to the sea floor can form complex ecological environments.

proximity to nature — the Sea of Japan, the Alps and natural wilderness are all within a short distance from the city centre.

The level of education here is very high, and people are proud of the fact that elementary and junior-high-school students from the area often rank first or second in national examinations. This is also reflected in Kanazawa's number of universities per capita, which is the third highest in Japan. People here like 'academic' things and appreciate the importance of science, so the atmosphere is fantastic both for conducting research and for living here as a researcher.

I think for foreigners, in particular, Kanazawa is a special place in Japan, where the old town and traditional customs such as tea ceremonies are very well preserved. Kanazawa City is one of the few major cities that were spared bombing during the Second World War, and so many traditions and much of the old architecture have survived.

I am researching the evolution of animals living in extreme environments, for example, near hydrothermal vents and cold seeps in the geologically active areas of the deep sea. The hypothesis I am working on is that such animals might have adapted from the communities that form near decaying organic matter on the sea floor, such as 'whale falls'. Last year, we reported on fossil sea-turtle remains from 80 million to 90 million years ago, long before the emergence of whales. I am lucky because Kanazawa University's marine station is located nearby on the Noto Peninsula, which, although just 50 kilometres or so from Kanazawa, is remote enough to allow us to study decaying processes. And for my work, the Japanese archipelago is very interesting, being one of the most geologically active areas in the world. I think research from Kanazawa will be able to shed light on the evolution of life on Earth. ■

INTERVIEWS BY BRETT DAVIS

Interviews have been edited for clarity and length.

MEET THE BIG PLAYERS

The three major research institutes in Kanazawa

Kanazawa Institute of Technology is a private higher-learning institute with more than 8,000 students and close to 350 teaching staff. It is known for its advanced mechanical-engineering facilities and workshops.

Japan Advanced Institute of Science and Technology is a postgraduate education and research institution. About 40% of its more than 1,100 students and 20% of its 150 faculty researchers are from overseas. Most foreign researchers obtain their positions by applying for an associate professor or five-year assistant professor position.

Kanazawa University is a prominent university on the Sea of Japan coast with more than 10,000 students and 1,000 teaching and faculty staff members. It has been active in increasing the number of international students and students studying abroad over the past two or three years, has developed collaborative and exchange relationships with dozens of universities worldwide and has introduced English-language programmes for students and faculty and staff members. B.D.